社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2009－CVIM－166 (29)
2009／3／14

頑健なヤコビ核主成分分析に向けて

藤木 淳[†] 赤穂 昭太郎[†]
日野 英逸[††] 村田 昇[††]

元来の核主成分分析では特徴空間におけるユークリッド計量に基づいた最小二乗推定によって主たる部分空間を求めるが，ヤコビ核主成分分析では入力空間におけるユークリッド計量に基づいて計算される重みを用いた重み付き最小二乗推定によって主たる部分空間を求める．一般に（重み付き）最小二乗推定量は外乱値に弱いため，本稿では RANSAC と χ^2 検定を利用して外乱値を検出することによりヤコビ核主成分分析の頑健化を行なう．

Towards robust Jacobian kernel PCA

Jun Fujiki,[†] Shotaro Akaho,[†] Hideitsu Hino[††]
and Noboru Murata[††]

Conventional kernel principal component analysis finds the principal subspace of the data by means of least squares estimation with respect to the metric defined in the feature space. on the other hand, Jacobian kernel principle component analysis (JKPCA) finds the principal subspace based on the metric in the input space by weighted least squares associated with the Jacobian of the feature map. For some applications such as image processing, it is more natural to utilize the metric in the input space, however, the weighted least squares estimation is sometimes too sensitive to outliers. To overcome this drawback of JKPCA and make it robust to outliers, an outlier detection method based on random sampling consensus (RANSAC) and χ^2 test is discussed, and its validity is confirmed by numerical experiments.

1. はじめに

観測データの**構造**の把握はデータ解析における基本的かつ重要な手順である．データの主構造の最も基本的な決定手法は**主成分分析**（**principal component analysis；PCA**）ある．PCA は，データを**ユークリッド**（**欧幾里得；欧氏**）空間の点とみなし，データの真値は欧氏空間内のアフィン空間上に分布するという仮定の下で有効な手法である．そして PCA ではデータと主構造とのずれである誤差を欧氏距離で測定し，誤差の二乗和を最小とする**最小二乗**（**least squares；LS**）基準を用いるが，これは誤差が等方正規分布に従うという仮定の下では最尤推定を与える．

しかしデータの真値や主構造がアフィン空間上に分布しない場合には主構造としてアフィン空間を仮定する PCA は不向きである．そこで**非線型主成分分析**（**nonlinear PCA；NLPCA**）が提案されてきた．NLPCA の考え方は，非線型構造をもつデータをアフィン空間上に分布させるために**特徴空間**（**feature space**）と呼ばれる高次元空間に**特徴写像**（**feature map**）と呼ばれる写像を行ない，その後に通常の PCA を行なう所にある．このとき，データの構造は特徴写像によってアフィン空間に写像される空間に限定されるため，データ構造の理解と特徴写像はほぼ等価である．よって特徴写像が理論的に既知の場合は問題がないが，未知の場合には，写像が先か構造が先かという鶏と卵の問題が生じる．しかしデータの構造を十分に近似できる写像が求まれば良いならば，多数の基底で表現される非常に高次元への写像を考えることにより鶏と卵の問題は近似的に解決される．

一般に特徴空間が高次元の場合，特徴空間における高次元ベクトルの計算量が多くなる．この問題を解決したのがカーネルトリック[1),21)]である．カーネルトリックとは，NLPCA に特徴写像の表現は不要で，特徴空間での誤差を測るための計量（内積）を与える対称な正定値カーネル関数があれば十分であるという議論のすりかえのことである．この議論のすりかえにより，計算量が特徴写像の次元数ではなくデータ数に依

† 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology
†† 早稲田大学 Waseda University

存することとなり非常に高い次元への特徴写像を考えても計算量が破綻しにくくなる．そしてまたガウシアンカーネルのような無限次元空間への写像を扱うこともできる．このカーネルトリックを用いたNLPCAを**核主成分分析（kernel PCA；KPCA）**[4),18)]と呼ぶ．

さて，KPCA[18)]は特徴空間における欧氏距離のLS基準によるあてはめ問題である．ゆえにあてはめ結果による入力データの推定値は特徴空間におけるあてはめ結果との最短距離であり，入力空間におけるあてはめ結果への最短距離とは限らない．そこでコンピュータビジョンでは入力空間における距離のLS基準によるあてはめ問題が研究されてきた．入力空間が欧氏空間の場合[2),5),14),15),17),20)]や球面の場合[9)]が提案され，そして一般的なリーマン空間の場合を核関数を用いて表現した**ヤコビ核主成分分析（Jacobian KPCA；JKPCA）**[10)]が提案された．

JKPCAは特徴写像を各観測データ点附近でJacobi行列を利用してアフィン写像に近似している．このアフィン写像により観測データ附近における入力空間と特徴空間の計量の対応を近似し，入力空間の誤差を特徴空間の誤差によって近似的に表現している．このことにより，入力空間のLS推定は特徴空間の重み付きLS推定により近似される．この近似は**等計量線（equi-metric curve）**，つまり入力空間においてデータから等距離にある閉曲線の特徴写像による像を超楕円体によって近似していることに相当する．しかし一般に特徴写像は非線型写像であるから，この近似は誤差が十分に小さい範囲内でしか成立せず，**外乱値（outlier）**，つまり誤差の大きいデータに対しては近似誤差と実際の誤差の乖離が大きく，外乱値を含む場合の推定結果は一般的に悪くなる．またLS推定自体モデルからデータの方向に過剰に適合し外乱値が推定に悪影響を及ぼすことが知られている．

そこで外乱値の影響を抑えるためにM推定量が提案された[13)]．M推定の基本的な考え方は，誤差分布が正規分布から少々ずれたとしても有効性がそれほど落ちないように外乱値の影響を制御した推定量を作る所にある．例えば平均の推定において誤差がラプラス分布に従うと仮定すると，最尤推定量は標本平均ではなく標本中央値となるというように，直感的には，誤差分布がM推定量から導かれる“正規分布関数と似た”確率分布（M推定量によっては正規化できず確率分布と見做せない場合もある）に従うと仮定した場合の最尤推定として捉えることができる．

本稿で提案する**頑健なヤコビ核主成分分析（Robust JKPCA；R-JKPCA）**は**ランザック（random sample consensus；RANSAC）**[6)]とχ^2検定の組合せによって外乱値を除去し，許容値のみで推定を行なう[19)]手法であるが，これも一種のM推定とみなせる（対応する確率分布は存在しない）．

提案手法であるR-JKPCAの有効性は，人工データを用いた実験によって評価される．

2. ヤコビ核主成分分析

入力空間であるm次元空間$\mathcal{R}$の点列に部分空間をあてはめる際，この点列を特徴空間であるn次元ヒルベルト空間$\mathcal{H}$に写像し，$\mathcal{H}$における線型あてはめに帰着することが多い．このあてはめを$\mathcal{H}$におけるLS基準で推定するのがKPCAである．しかし入力空間に自然な計量が定義されている場合に特徴空間における計量を基準として推定を行うことは好ましくない[3)]ため，入力空間の計量における近似LS基準で部分空間をあてはめるJKPCA[10)]が提案された．

2.1 入力空間の計量と特徴空間の計量

入力空間をm次元リーマン空間$\mathcal{R}$とし，$\mathcal{R}$上の点$\boldsymbol{x}$におけるリーマン計量を$G_{\boldsymbol{x}}$とする．また，観測データ点を$\{\boldsymbol{x}_{[d]}\}_{d=1}^{D}$とする．JKPCAでは観測データ点$\boldsymbol{x}_{[d]}$附近の空間を計量が$G_{\boldsymbol{x}_{[d]}}$で一定であるアフィン空間で近似する．つまり$\boldsymbol{x}=\boldsymbol{x}_{[d]}+\delta\boldsymbol{x}$なる点$\boldsymbol{x}$と観測データ点$\boldsymbol{x}_{[d]}$の距離$r^p$を$(r^p)^2=(\delta\boldsymbol{x})^\top G_{\boldsymbol{x}_{[d]}}(\delta\boldsymbol{x})$で近似する．点$\boldsymbol{x}$を$n$次元ヒルベルト空間である特徴空間$\mathcal{H}$に特徴写像$\boldsymbol{\phi}:\boldsymbol{x}\mapsto\boldsymbol{\phi}(\boldsymbol{x})$で射影する．ここで特徴写像$\boldsymbol{\phi}$のヤコビ行列$J_{\boldsymbol{\phi}}$を用いると，特徴写像は観測データ附近で

$$\boldsymbol{x}\mapsto\boldsymbol{\phi}_{[d]}+J_{\boldsymbol{\phi}_{[d]}}(\boldsymbol{x}-\boldsymbol{x}_{[d]})$$

$(\boldsymbol{\phi}_{[d]}=\boldsymbol{\phi}(\boldsymbol{x}_{[d]}))$と近似でき，

$$(r^p)^2=(\delta\boldsymbol{\phi}(\boldsymbol{x}))^\top\mathcal{G}_{\boldsymbol{\phi}_{[d]}}\delta\boldsymbol{\phi}(\boldsymbol{x})$$

$$\left[\begin{array}{l}\boldsymbol{\phi}(\boldsymbol{x}+\delta\boldsymbol{x})=\boldsymbol{\phi}(\boldsymbol{x})+\delta\boldsymbol{\phi}\\ \mathcal{G}_{\boldsymbol{\phi}_{[d]}}=(J^+_{\boldsymbol{\phi}_{[d]}})^\top G_{\boldsymbol{x}_{[d]}}J^+_{\boldsymbol{\phi}_{[d]}}\\ \mathcal{G}^{-1}_{\boldsymbol{\phi}_{[d]}}=J_{\boldsymbol{\phi}_{[d]}}G^{-1}_{\boldsymbol{x}_{[d]}}J^\top_{\boldsymbol{\phi}_{[d]}}\end{array}\right],$$

（X^+はXのムーア・ペンローズ逆行列）と近似できる．

2.2 特徴空間における線型あてはめ

特徴空間のデータ$\boldsymbol{\phi}_{[d]}=\boldsymbol{\phi}(\boldsymbol{x}_{[d]})$に対して$n-1$次元アフィン空間$\boldsymbol{a}^\top\boldsymbol{\phi}+b=0$をあてはめるには，

$$\mathcal{E}(\boldsymbol{a})=\sum_{d=1}^{D}\frac{(\boldsymbol{a}^\top\boldsymbol{\phi}_{[d]}+b)^2}{\boldsymbol{a}^\top\mathcal{G}^{-1}_{\boldsymbol{\phi}_{[d]}}\boldsymbol{a}}\qquad(1)$$

を最小にする$\boldsymbol{a}$，bを求めれば良い[10)]．ここでは

$$\widetilde{\boldsymbol{\phi}} = \begin{pmatrix} \boldsymbol{\phi} \\ 1 \end{pmatrix}, \quad \widetilde{\boldsymbol{a}} = \begin{pmatrix} \boldsymbol{a} \\ b \end{pmatrix}, \quad \widetilde{\mathcal{G}}_{\boldsymbol{\phi}_{[d]}} = \begin{pmatrix} \mathcal{G}_{\boldsymbol{\phi}_{[d]}} & \mathbf{0} \\ \mathbf{0}^\top & 0 \end{pmatrix}$$

と置いて $n+1$ 次元空間の原点を通る n 次元線型部分空間 $\widetilde{\boldsymbol{a}}^\top \widetilde{\boldsymbol{\phi}} = 0$ の推定問題に帰着させる，つまり

$$\mathcal{E}(\widetilde{\boldsymbol{a}}) = \sum_{d=1}^{D} \frac{\widetilde{\boldsymbol{a}}^\top \left[\widetilde{\boldsymbol{\phi}}_{[d]} \widetilde{\boldsymbol{\phi}}_{[d]}^\top\right] \widetilde{\boldsymbol{a}}}{\widetilde{\boldsymbol{a}}^\top \widetilde{\mathcal{G}}^{+}_{\boldsymbol{\phi}_{[d]}} \widetilde{\boldsymbol{a}}} \tag{2}$$

の最小にする $\widetilde{\boldsymbol{a}}$ を求めれば良く[2),14),20)]，式 (2) の最小化を核関数を用いて行なう．

2.3 核関数による表現

本稿では，核関数として微分可能なものを考え，核関数 $k(\boldsymbol{x}, \boldsymbol{y}) = \widetilde{\boldsymbol{\phi}}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\phi}}(\boldsymbol{y})$ だけでなく，その微分

$$\underset{(m\times 1)}{\boldsymbol{k}(\boldsymbol{x}, \boldsymbol{y})} = \frac{\partial k(\boldsymbol{x}, \boldsymbol{y})}{\partial \boldsymbol{x}^\top} = J_{\widetilde{\boldsymbol{\phi}}}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\phi}}(\boldsymbol{y})$$

も用いる．これを**ヤコビ核 (Jacobian kernel)** と呼ぶ．また $\boldsymbol{a}$ の存在範囲として $\widetilde{\boldsymbol{\phi}}_{[d]}$ の線型結合

$$\underset{(n\times 1)}{\boldsymbol{a}} = \sum_{p} \alpha_{[d]} \widetilde{\boldsymbol{\phi}}_{[d]} = \underset{(n\times D)}{\widetilde{\boldsymbol{\Phi}}} \underset{(D\times 1)}{\boldsymbol{\alpha}}$$

だけを考える．ここで $D \times D$ 行列 $\mathcal{K}$ を

$$(\mathcal{K})_{ij} = k(\boldsymbol{x}_{[i]}, \boldsymbol{x}_{[j]}),$$
$$\mathcal{K} = \begin{pmatrix} \mathcal{K}_{[1]} & \cdots & \mathcal{K}_{[D]} \end{pmatrix}$$

で定義し，また，$D \times m$ 行列 $\boldsymbol{\mathcal{K}}_{[d]}$ を

$$\boldsymbol{k}_{[i][j]} = \boldsymbol{k}(\boldsymbol{x}_{[i]}, \boldsymbol{x}_{[j]}),$$
$$\boldsymbol{\mathcal{K}}_{[d]} = \begin{pmatrix} \boldsymbol{k}_{[d][1]} & \cdots & \boldsymbol{k}_{[d][D]} \end{pmatrix}^\top$$

で定義すると，

$$\mathcal{K}_{[d]} = \widetilde{\boldsymbol{\Phi}}^\top \widetilde{\boldsymbol{\phi}}_{[d]}, \quad \boldsymbol{\mathcal{K}}_{[d]} = \widetilde{\boldsymbol{\Phi}}^\top J_{\widetilde{\boldsymbol{\phi}}_{[d]}}$$

が成立し，このとき式 (2) は，写像 $\boldsymbol{\phi}$ を含まない

$$\mathcal{E}'(\boldsymbol{\alpha}) = \sum_{d=1}^{D} \frac{\boldsymbol{\alpha}^\top \mathcal{K}_{[d]} \mathcal{K}_{[d]}^\top \boldsymbol{\alpha}}{\boldsymbol{\alpha}^\top \boldsymbol{\mathcal{K}}_{[d]} \widetilde{\mathcal{G}}^{+}_{\boldsymbol{x}_{[d]}} \boldsymbol{\mathcal{K}}_{[d]}^\top \boldsymbol{\alpha}} \tag{3}$$

となり，これを最小にする $\boldsymbol{\alpha}$ を求めれば良い．ここで

$$\widetilde{\mathcal{G}}_{\boldsymbol{x}_{[d]}} = \begin{pmatrix} \mathcal{G}_{\boldsymbol{x}_{[d]}} & \mathbf{0} \\ \mathbf{0}^\top & 0 \end{pmatrix}$$

である．

2.4 解法とアルゴリズム

式 (3) の最小化アルゴリズム[2)] を紹介する．$\boldsymbol{\alpha}$ の近似値 $\widehat{\boldsymbol{\alpha}}$ が得られたとき，

$$\mu_{[d]} = \widehat{\boldsymbol{\alpha}}^\top \boldsymbol{\mathcal{K}}_{[d]} \widetilde{\mathcal{G}}^{+}_{\boldsymbol{x}_{[d]}} \boldsymbol{\mathcal{K}}_{[d]}^\top \widehat{\boldsymbol{\alpha}},$$
$$\Lambda = \mathrm{diag}\{\mu_{[1]}, \ldots, \mu_{[D]}\}$$

とすると $\mathcal{E}'(\boldsymbol{\alpha}) \approx \boldsymbol{\alpha}^\top \mathcal{K} \Lambda^{-1} \mathcal{K} \boldsymbol{\alpha}$ と近似できるので，

$$\boldsymbol{\alpha} = \mathtt{UnitMinEigenVec}\left[\mathcal{K}\Lambda^{-1}\mathcal{K}\right]$$

($\mathtt{UnitMinEigenVec}[X]$ は正方行列 X の最小固有値に対応する単位固有ベクトル) となる．

以下のアルゴリズムにおいて右肩の $[k]$ によって k ステップ目の値を表すものとする．初期値は右肩が [0] となる（初期値の設定は後述）．

(1) 初期値 $\mu_{[d]}^{[0]}$ $(d = 1, \ldots, D)$ の設定

(2) 収束するまで (a)，(b) を繰り返す

(a) $\widehat{\boldsymbol{\alpha}}^{[k+1]} = \mathtt{UnitMinEigenVec}[\mathcal{K}(\Lambda^{[k]})^{-1}\mathcal{K}]$

(b) $\mu_{[d]}^{[k+1]} = (\widehat{\boldsymbol{\alpha}}^{[k+1]})^\top \boldsymbol{\mathcal{K}}_{[d]} G^{+}_{\boldsymbol{x}_{[d]}} \boldsymbol{\mathcal{K}}_{[d]}^\top (\widehat{\boldsymbol{\alpha}}^{[k+1]})$ で $\{\mu_{[d]}\}_{d=1}^{D}$ を更新．

2.5 初 期 値

上記アルゴリズムの初期値の設定方法として以下のような方法がある．

2.5.1 核主成分分析 (KPCA)

$\{\mu_{[d]}^{[0]} = 1\}_{d=1}^{D}$ の場合[2)] で

$$\mathcal{E}'(\boldsymbol{\alpha}) \approx \boldsymbol{\alpha}^\top \mathcal{K}^2 \boldsymbol{\alpha}$$

を最小化する $\widehat{\boldsymbol{\alpha}} = \mathtt{UnitMinEigenVec}[\mathcal{K}^2]$ を初期値とする．

2.5.2 欧氏化 (Euclideanization)

欧氏化[7),10),11)] とは特徴写像によるデータ点附近の線分長の平均拡大率の逆数をデータ点に重みとして与える手法であり，

$$\mathcal{E}'(\boldsymbol{\alpha}) \approx \boldsymbol{\alpha}^\top \left(\mathcal{K} D^{\frac{1}{m}} \mathcal{K}\right) \boldsymbol{\alpha},$$
$$D = \mathrm{diag}\left\{\mathrm{Det}\, \mathcal{G}_{\boldsymbol{\phi}_{[1]}}, \ldots, \mathrm{Det}\, \mathcal{G}_{\boldsymbol{\phi}_{[D]}}\right\}$$
$$\mathrm{Det}\, \mathcal{G}_{\boldsymbol{\phi}_{[d]}} = \det\left\{J^{+}_{\boldsymbol{\phi}_{[d]}} (J^{+}_{\boldsymbol{\phi}_{[d]}})^\top\right\} \cdot \det G_{\boldsymbol{x}_{[d]}}$$

($\mathrm{Det}\, X$ で X の 0 でない固有値の積を表す) を最小化する $\widehat{\boldsymbol{\alpha}} = \mathtt{UnitMinEigenVec}[\mathcal{K} D^{\frac{1}{m}} \mathcal{K}]$ を初期値とする．

2.5.3 Taubin 法[20)]

各データ点で異なる計量を核空間で平均した $\mathrm{E}[G_{[d]}^{\mathrm{kernel}}] = \sum_{d=1}^{D} \boldsymbol{\mathcal{K}}_{[d]} G^{+}_{\boldsymbol{x}_{[d]}} \boldsymbol{\mathcal{K}}_{[d]}^\top$ で近似する手法であり，

$$\mathcal{E}'(\boldsymbol{\alpha}) \approx \frac{\boldsymbol{\alpha}^\top \mathcal{K}^2 \boldsymbol{\alpha}}{\boldsymbol{\alpha}^\top \left(\mathrm{E}[G_{[d]}^{\mathrm{kernel}}]\right) \boldsymbol{\alpha}}$$

を最小化する $\widehat{\boldsymbol{\alpha}}$，つまり一般化固有値問題

$$\mathcal{K}^2 \boldsymbol{\alpha} = \lambda \left(\mathrm{E}[G_{[d]}^{\mathrm{kernel}}]\right) \boldsymbol{\alpha}$$

の最小固有値に対応する固有ベクトルを初期値とする．

2.6 ヤコビ核主成分分析

KPCA は $\mathcal{K}^2 = \mathcal{K}\mathcal{K}$ の固有ベクトルを固有値の大きい順に並べて新しい座標枠を作る手法であり，JKPCA

は $\mathcal{K}(\Lambda^{[\infty]})^{-1}\mathcal{K}$ の固有ベクトルを固有値の大きい順に並べて新しい座標枠を作る手法である．よって特徴ベクトル $\widetilde{\boldsymbol{\phi}}_{[d]}$ に対して重み

$$(\mu_{[d]}^{[\infty]})^{-\frac{1}{2}} = ||\boldsymbol{\mathcal{K}}_{[d]}^{\top}\boldsymbol{\alpha}||_{G^{+}_{\boldsymbol{x}_{[d]}}}^{-1}$$

を与えたときの重みつき PCA である．この重みはデータ点 $\boldsymbol{x}_{[d]}$ 附近の計量と $\widetilde{\boldsymbol{\phi}}_{[d]}$ 附近の計量を比べ，$\widetilde{\boldsymbol{\phi}}_{[d]}$ 附近に局所的に入力空間の計量を反映させたものである．

3. ヤコビ核主成分分析の頑健化

3.1 入力空間の近似誤差

JKPCA は入力空間における誤差 $r_{[d]}$ を

$$r_{[d]}^2 \approx \frac{\boldsymbol{\alpha}^{\top}\mathcal{K}_{[d]}\mathcal{K}_{[d]}^{\top}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^{\top}\boldsymbol{\mathcal{K}}_{[d]}\widetilde{G}^{+}_{\boldsymbol{x}_{[d]}}\boldsymbol{\mathcal{K}}_{[d]}^{\top}\boldsymbol{\alpha}} \tag{4}$$

（$\boldsymbol{\alpha}$ は JKPCA の収束時の値）と近似しており，この近似誤差

$$R_{[d]} = \sqrt{\frac{\boldsymbol{\alpha}^{\top}\mathcal{K}_{[d]}\mathcal{K}_{[d]}^{\top}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^{\top}\boldsymbol{\mathcal{K}}_{[d]}\widetilde{G}^{+}_{\boldsymbol{x}_{[d]}}\boldsymbol{\mathcal{K}}_{[d]}^{\top}\boldsymbol{\alpha}}} \tag{5}$$

を判断基準として JKPCA を頑健化する．

3.2 RANSAC

RANSAC[6] は，ランダムサンプリングに基づくロバストなモデル作成手法であり，

(1) 部分空間を形成する最小のデータ数を用いて部分空間をあてはめる．

(2) あてはめ結果に対して，全データの各々が許容値かどうかを検定し，許容値の個数を数える．

(3) ランダムランプリングを一定回数繰り返し，許容値の個数が最大となるものを選ぶ．

(4) 選ばれた個数が最大となる部分空間を用いて全データの各々が例外値かどうかを検定し，例外値を取り除く．

(5) 例外値を除いたデータ全てから部分空間をあてはめる．

という手続きによって部分空間をあてはめる手法である．

データ量が十分にあるとし，汚染率（外乱値の割合）が α，部分空間を作成するのに必要なデータ数を m 個であるとする．このとき m 個のデータが全て許容値である確率が β 以上となるためのランダムサンプリングの回数 k は

$$k \geq \frac{\log_{10}(1-\beta)}{\log_{10}\{1-(1-\alpha)^m\}}\ . \tag{6}$$

をみたす．

3.3 MAD 推定量

中央絶対偏差（median absolute deviation；MAD）とは，1 次元実データ $\{x_n\}_{n=1}^N$ に対して

$$\mathrm{mad}\,[x_n] = \mathrm{med}\left[\,|x_n - \mathrm{med}\,(x_n)|\,\right]$$

によって定義される統計量である．標準偏差の定義

$$\sigma = \sqrt{\mathrm{Var}\,[x_n]} = \sqrt{\mathrm{E}\,[\,(x_n - \mathrm{E}[x_n]\,)^2]}$$

と比較しみてればわかるように，MAD は標準偏差と同様，確率分布の尺度助変数となる．

ここで $x_n \sim \mathrm{N}(0,\sigma^2)$ のとき，$N \to \infty$ における漸近的な振舞いを考えると，

$$\mathrm{E}[x_n] \simeq 0, \quad \mathrm{med}\,[x_n] \simeq 0$$

が成立するので，

$$\mathrm{mad}\,[x_n] \simeq \mathrm{med}\left[\,|x_n|\,\right], \quad \sigma^2 \simeq \mathrm{E}[x_n^2]$$

となる．一方

$$\frac{x_n^2}{\sigma^2} \sim \chi^2(1)$$

であるから，

$$\mathrm{med}\,\left[x_n^2\right] \simeq \chi^2_{0.5}(1)\cdot\sigma^2$$

となり，

$$\sigma \simeq \frac{1}{\sqrt{\chi^2_{0.5}(1)}}\mathrm{mad}\,[x_n] \approx 1.4826\,\mathrm{med}\left[\,|x_n|\,\right]$$

となる[16]．ここで $\chi^2_{\alpha}(L)$ は自由度 L の χ^2 分布の $100\alpha\%$ 点である．

これと同様に，L 次元データ $\{\boldsymbol{x}_n\}_{n=1}^N$ が等方正規分布 $\boldsymbol{x}_n \sim \mathrm{N}(\mathbf{0}_L, \sigma^2\mathrm{I}_L)$ に従うと仮定すると，

$$\frac{||\boldsymbol{x}_n||^2}{\sigma^2} \sim \chi^2(L)$$

であるから，漸近的に

$$\sigma \simeq \frac{1}{\sqrt{\chi^2_{0.5}(L)}}\mathrm{med}\left[||\boldsymbol{x}_n||\right] \tag{7}$$

が成立する．

本稿では m 次元入力空間を n 次元特徴空間に射影した後に n 次元特徴空間における k 次元あてはめ問題を考える．このときモデルからのずれである誤差ベクトルは $L = n - k$ 次元空間のベクトルとなる．しかし，入力空間の像は特徴空間内の m 次元以下の図形へと射影されるため，誤差ベクトルの分布には何らかの偏りが生ずる．しかし問題の単純化のため，誤差ベクトルの分布が等方正規分布 $\mathrm{N}(\mathbf{0}_L, \sigma^2\mathrm{I}_L)$ に従うと仮定して，mad 推定量と σ の関係を利用し，誤差ベクトルのノルムの分散を推定する．

3.4 RANSAC と χ^2 検定

RANSAC と χ^2 検定による外乱値除去手法[19] を JKPCA に以下のように適用する．本稿では，検定に用いる誤差分散の推定量 $\widehat{\sigma}^2$ の定め方も与える．

3.4.1 閾値となる誤差分散の決定

(1) RANSAC で部分空間のあてはめて残差の中央値を計算する，という試行を式 (6) で定まる回数行なう．

(2) 全ての試行のうち中央値が最小となるものに対して式 (7) から誤差分散の推定量 $\widehat{\sigma}$ を求める．

3.4.2 Robust JKPCA (R-JKPCA)

(1) n 次元特徴ベクトルの集合 $\{\boldsymbol{\Phi}(\boldsymbol{x}_{[d]})\}_{d=1}^{D}$ からランダムに $\{\boldsymbol{\Phi}(\boldsymbol{y}_{[m]})\}_{m=1}$ を選び JKPCA により $\widetilde{\boldsymbol{\alpha}}$ を求め，全データに対して式 (5) の近似残差 $R_{[d]}$ を計算し，$R_{[d]}^2 < \widehat{\sigma}^2$ なる個数を S とする．

(2) **(1)** を反復して S を最大とする主成分部分空間を選び，近似残差を有意水準 $100(1-\gamma)\%$ で検定，つまり $R_{[d]}^2 \geq \chi_{\gamma}^2(L)\cdot\widehat{\sigma}^2$ をみたすデータを外乱値として除去する．

(3) 許容値から JKPCA で主成分部分空間を推定する．

4. 実　　験

4.1 2 次曲線のあてはめ

2 次曲線のあてはめ問題を用いて提案手法の有効性を示す．

許容値は放物線 $y=x^2$ 上の点を x 座標が一様分布 $U[-3,3]$ に従うように生成し，各座標に**切断正規分布 (truncated normal distribution)** ☆

$$\mathrm{N}_{[-2.32\sigma,2.32\sigma]}(0,\sigma^2)$$

($\sigma=0.2$) から生成した誤差を添加したもの，外乱値は長方形内の一様分布 $\mathrm{U}[[-6,6]\times[-1.5,10.5]]$ から生成したデータ $(x_{[d]},y_{[d]})$ のうち，$y=x^2$ との近似距離 $R_{[d]}$ が 3σ 以上のものとした．

汚染率を $\alpha\%$ とし，許容値を $100(1-\alpha)$ 点，外乱値を 100α 点（合計 100 点）生成し，2 次曲線のあてはめを行なう．

2 次曲線は平面上の 5 点によって決定されるので本稿の実験では $m=5$ であり，$\beta=0.99$ とすると，式 (6) により汚染率 $\alpha=0.462\cdots$ のときに $k=100$ となるので，RANSAC におけるランダムサンプル数は 100 とする．また近似誤差は有意水準 1% ($\gamma=0.99$) で検定し，各汚染率に対して 100 回のあてはめを行なって統計的処理を行なった．

4.2 外乱値の検出精度

データが許容値であることを帰無仮説，外乱値であることを対立仮説とする．このとき，許容値を外乱値と判定することを**第一種過誤**，外乱値を許容値と判定することを**第二種過誤**と呼ぶ．

表 1 に汚染率（外乱値の割合）に対する 2 つの過誤の割合を**平均値 ± 標準偏差**の形で記載した．表 1 により，汚染率が 40% 以下でなら 9 割以上のデータが正しく分類できている．

表 1 汚染率と過誤の割合 (%)：100 点

汚染率	第一種過誤	第二種過誤
0%	2.34 ± 1.66	−
10%	3.73 ± 3.08	2.80 ± 5.87
20%	4.94 ± 4.37	3.40 ± 5.12
30%	6.19 ± 5.33	3.73 ± 4.06
40%	7.78 ± 6.57	5.60 ± 5.51
50%	24.82 ± 11.72	15.10 ± 8.69

4.3 あてはめ結果

図 1 は汚染率 10% のデータへのあてはめ結果である．図 2 は汚染率 30%，50% の場合の R-JKPCA によるあてはめ結果である．これら図において黒丸は許容値として生成したデータであり，罰点は例外値として生成したデータである．また点線は $y=x^2$ であり，実線はあてはめ結果である．図 1 より JKPCA は外乱値の影響で悪い結果だが，R-JKPCA は外乱値の影響を軽減していることがわかり，図 2 より汚染率が 50% となると計算が破綻することもあるが，概ねうまくあてはまっていることがわかる．

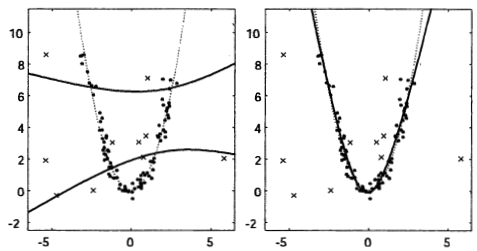

図 1 JKPCA (左) と R-JKPCA (右) (100 点，汚染率 10%)

4.4 許容値に対する近似誤差と真の誤差

あてはめにおいて，検定により許容値と判定されたデータに対し，近似誤差 $R_{[d]}$ 及びデータと $y=x^2$ の距離（真の距離と呼ぶ）が汚染率に対してどれだけ頑健か調べた．図 3 は近似誤差と真の誤差の平均と分散

☆ 平均 μ，分散 σ^2 の正規分布のうち区間 $[A,B]$ の部分を正規化した切断正規分布を $\mathrm{N}_{[A,B]}(\mu,\sigma^2)$ で表すこととする．この切断正規分布の密度関数は $A<x<B$ において

$$\frac{\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)}{\Phi\left(\frac{B-\mu}{\sigma}\right)-\Phi\left(\frac{A-\mu}{\sigma}\right)}$$

となる．ここで $\phi(x)$ 及び $\Phi(x)$ は標準正規分布の密度関数及び累積分布関数である．

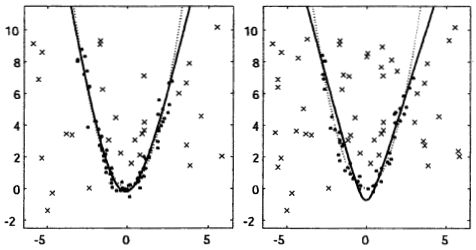

図 2 R-JKPCA；汚染率 30%（左），50%（右）

である．実線，点線はそれぞれ近似誤差及び真の誤差の平均であり，左及び右の縦棒は近似距離（左）及び真の距離（右）それぞれの標準偏差を表す．実験により，汚染率の増加に比べて許容値一個あたりの誤差が抑えられており，提案手法の頑健性が確認できた．

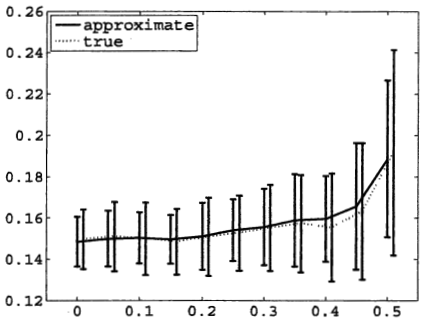

図 3 汚染率と許容値あたりの近似誤差

参 考 文 献

1) M. Aizerman, É. Braverman and L. Rozonoér, "Theoretical foundations of the potential function method in pattern regognition learning," Automation and Remote Control, **25**:821-837, 1964.
2) S. Akaho, "Curve fitting that minimizes the mean square of perpendicular distances from sample points," SPIE, Vision Geometry II, 1993.
3) 赤穂昭太郎, "入力空間でのマージンを最大化するサポートベクタマシン," 信学論 D-II, **J86-D-II**(7):934-942, 2003.
4) 赤穂昭太郎, "カーネル多変量解析-非線形データ解析の新しい展開-," 岩波書店, 2008.
5) W. Chojnacki, M. J. Brooks, A. van den Hangel and D. Gawley, "On the fitting of surface to data with covariances," IEEE TPAMI, **22**(11):1294-1303, 2000.
6) M. A. Fischer and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," Comm. ACM,**24**:381-395, 1981
7) 藤木淳, 赤穂昭太郎, "球面上の点列への小円あてはめ〜カメラ運動の平滑化に向けて〜," 信学技報 PRMU2004-149:91-96, 2004(12).
8) J. Fujiki and S. Akaho, "Spherical PCA with Euclideanization," Subspace 2007(ACCV07).
9) 藤木淳, 赤穂昭太郎, "球面最小二乗法による球面上の曲線あてはめ," Subspace 2008(MIRU2008).
10) 藤木淳, 赤穂昭太郎, 入力空間での計量に基づいた核主成分分析, 信学技報, **108**(327):69-74, 2008.
11) 藤木淳, 赤穂昭太郎, 日野英逸, 村田昇, 主成分曲線のあてはめによる放射対称歪曲の較正, 信学技報, **108**(363):13-18, 2008.
12) R. Hartley and A. Zisserman. *Multiple view geometry in computer vision.* Cambridge University, Cambridge, 2nd edition, 2003.
13) P. J. Huber, "Robust estimation of a location parameter," Ann. Math. Stat., **35**:73-101, 1964.
14) 金谷健一, 菅谷保之, "幾何学的当てはめの厳密な最尤推定の統一的計算法," 情処研報, 2008-CVIM-164-3:17-24, 2008.
15) Y. Leeden and P. Meer, "Heteroscedastic regression in computer vision: problems with bilinear constraint," IJCV, **37**(2):127-150, 2000.
16) R. J. Rousseeuw and A. M. Leroy, Robust Regression and Outlier Detection, John Wiley & Sons, NY, 1987.
17) P.D.Sampson, "Fitting conic sections to 'very scattered' data: an iterative refinement of the Bookstein algorithm," Comput. Vision, Graphics, and Image Processing, **18**:97-108, 1982.
18) B.Schölkopf, A.Smola and K.-R.Müller, "Nonlinear component analysis as a kernel eigenvalue problem," Neural Computation, **10**:1299-1319, 1998.
19) Y. Sugaya and K. Kanatani, "Outlier removal for motion tracking by subspace separation," IEICE Trans. Inf.&Syst., **E86-D**(6):1095-1102, 2003.
20) G. Taubin, "Estimation of planar curves, surfaces, and nonplanar space curves defined by implicit equations with applicatons to edge and range image segmentation," IEEE TPAMI, **13**(11):1115-1138, 1991.
21) V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, 1995.