様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月 ２０ 日現在

機関番号：３２６４３
研究種目：若手研究
研究期間：2019～2022
課題番号：１９Ｋ２０２２４

研究課題名（和文）異常度そのものへのスパースモデリングの新規適用の検討

研究課題名（英文）Investigation of a new application of sparse modeling to an abonormality itself

研究代表者
小林　靖之（Kobayashi, Yasuyuki）

帝京大学・理工学部・准教授

研究者番号：００６０４５１３

交付決定額（研究期間全体）：（直接経費）　　800,000円

研究成果の概要（和文）：スパース化した標本マハラノビス距離モデルとして、未知データxと学習データの標本共分散行列Sに基づく一次連立方程式をCoordinateDescent法で解いて得たStudent化主成分ベクトルを用いるモデルを提案した。
母固有値が0値の標本主成分と正値の母固有値の標本主成分との差異を明らかにする必要があるため、正値の母固有値の標本主成分について標本固有値・ベクトルのバラツキを考慮した標本Student化主成分モデルを提案した。

研究成果の学術的意義や社会的意義
標本マハラノビス距離(SMD)やRidge正則化SMDでは数値計算上の誤差により0値を中心に広がった分布をもつが、提案したスパース化SMDでは正確に0値のみをもつため、SMDの数値計算上の誤差に由来する不安定現象を除去可能となると期待される。
正値の母固有値の標本主成分について標本固有値・ベクトルのバラツキを考慮した標本Student化主成分モデルを用いれば、SMDの大きい試験データについて個々のスチューデント化主成分要素を検定した要因分析が可能になる。

研究成果の概要（英文）：For a model of sparse sample Mahalanobis distance, we proposed a model using a studentized principal component vector calculated by Coordinate Descent method to solve simultaneous linear equations including unknown vector x and sample covariance matrix S for learning data.
To show clearly a difference of sample principal components corresponding population eigenvalue of zero and of positive value, we proposed a model of sample studentized principal component considering the fluctuation of sample eigenvalues and vectors for population eigenvalue of positive value.

研究分野：統計科学

キーワード：スパースモデリング　異常度に対するスパース化　マハラノビス距離　数値計算上の安定条件　正則化係数

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

機械学習で問題となる過学習防止や変数の意味づけ困難を解決するため、多くの変数をもつデータを少ない変数で表すスパースモデリングが注目を集めている。教師付き学習では大きな成果を挙げているが、半教師付き学習である未知異常サンプルを検出する目的の異常度へのスパース化は学習サンプルだけに留まり、異常度自体はスパース化されていなかった。

２．研究の目的

本研究では、学習サンプルではなく異常度モデル自体をスパース化したモデルをマハラノビス距離に対して提案し、その数値計算上の安定条件の提案・証明や、スパース化に必要な正則化係数を数値実験無しでの決定法を提案し、その妥当性を明らかにする。つまり、スパース化した標本マハラノビス距離モデルとその数値安定条件の提案・証明や、スパース化した異常度における数値実験不要な正則化係数ρの決定法の提案を行なう。

３．研究の方法

本研究では、スパース化した標本マハラノビス距離モデルや、一般的な正値の母固有値の標本主成分について数理モデルを導出した後、モンテカルロシミュレーションにより標本マハラノビス距離や標本主成分の標本分布を得て上記数理モデルの妥当性を検討した。

４．研究成果

スパース化した標本マハラノビス距離モデルとその数値安定条件の提案・証明のため、下記(1)のモデルを提案した。スパース化した標本マハラノビスのモデルについて数値安定条件を求めるに当り、スパース化で除かれる母固有値 0 の標本主成分は計算機の浮動小数点演算による丸め誤差の影響を受けるため、この影響を受けた母固有値 0 の標本主成分モデルの検討を続けた。その過程で、母固有値が 0 値の標本主成分モデルが正値の母固有値の標本主成分とどのように異なるのか明らかにする必要があるため、0 値ではなく一般的な正値の母固有値の標本主成分について標本固有値・ベクトルのバラツキを考慮した、下記(2)の標本主成分モデルを提案した。ただし、母固有値が 0 値の標本主成分モデルや数値安定条件の提案には至らなかった。

また、スパース化した異常度における数値実験不要な正則化係数ρの決定法の提案のため、母固有値 0 の標本主成分は正確に 0 であるが計算機の浮動小数点演算で生じる数値誤差によって微小な正値になる事実から、浮動小数点演算による丸め誤差の影響を受けた母固有値 0 が微小正値の標本固有値になり②のρ値の候補となるため、標本共分散行列計算における丸め誤差伝搬のモデル化を進め、母固有値 0 の標本固有値の分布モデルを検討したが論文投稿に至らなかった。等しい母固有値が異なる標本固有値を取る現象をモデル化するためにランダム行列理論と順序統計学の知見を応用した近似モデルを検討した論文は却下されたので、母固有値が若干異なる場合の近似モデルも含めて証明過程を再検討している。

(1) スパース化した標本マハラノビス距離モデル [1]

統計的機械学習の１手法として、標本マハラノビス距離（Sample Mahalanobis distance; SMD）は多変数で状態が表されるプラント等の異常判定のために産業で広く用いられる。SMD は学習データxの基準分布(平均$\bar{x}$)の従う多変量正規分布の確率密度関数の指数に相当し、状態を示すベクトルyの発生頻度が小さい（異常である）ほど SMD は大きい。

$$\mathrm{pdf}(y) = \frac{1}{\sqrt{(2\pi)^p|\mathrm{S}_x|}} e^{-\frac{(y-\bar{x})'\mathrm{S}_{\bar{x}}^{-1}(y-\bar{x})}{2}} = \frac{1}{\sqrt{(2\pi)^p|\mathrm{S}_x|}} e^{-\frac{1}{2}\sum_{i=1}^{p}\frac{\{(y-\bar{x})\cdot f_i\}^2}{l_i}} = \frac{1}{\sqrt{(2\pi)^p|\mathrm{S}_x|}} e^{-\frac{\mathrm{SMD}}{2}}$$

状態yの発生頻度は基準分布の確率密度に比例するが、基準分布が球状ではない場合（変数間に相関がある場合）、基準分布の中心からのユークリッド距離は基準分布の等高線と一致しない。そこで、変数間の相関を無くすために、基準分布が学習した共分散行列S_xの固有ベクトルf_iを軸とする超楕円体で表せるので、固有ベクトルf_iの軸長を対応する固有値l_iで規格化した座標系によるユークリッド距離により SMD を得る。しかし、学習データの真の共分散行列Σ_xの固有値λ_iが実際には 0 である場合、数値誤差で非常に小さい正数となったl_iによる不正確な規格化により、SMD の挙動は不安定化する。

少数の学習データから本質を抽出する意図で、少ない変数で表せる構造をデータがもつと仮定する、スパースモデリングが注目を集めている。スパースモデリングの例として、回帰分析において LASSO (Least Absolute Selection and Shrinkage Operator)が挙げられる。回帰分析のモデル式 $y = X\beta$（目的変数ベクトルy、説明変数のデータ行列X、偏回帰係数ベクトルβ)について、$\|y - X\beta\|_2^2$ を最小化するβを推定解とするが、LASSO ではℓ_1ペナルティを加えた$\|y - X\beta\|_2^2$を最小化するβを推定解とする。適切な正則化係数ρを与えると、βの少

数要素以外が厳密に 0 値となり、 βの変数選択や解釈が可能となる。

$$\underset{\beta}{\operatorname{argmin}}\{\|y - X\beta\|_2^2 + \rho\|\beta\|_1\} = \underset{\beta}{\operatorname{argmin}}\left\{\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{p} X_{ij}\beta_j\right)^2 + \rho\sum_{j=1}^{p}|\beta_j|\right\}$$

スパースモデリングは機械学習モデル中の変数選択の有力な処方箋である。また、スパースモデリングの異常検出への適用も検討され始め、マハラノビス距離のf_iに当たる主成分分析のローディングベクトルのスパース化や、Graphical LASSO による学習データの共分散行列のスパースモデルを用いて未知データの異常度を推定する手法が提案されている。

ところが、異常度のスパースモデリングの研究背景を振り返ると、学習データの共分散行列に対するスパース化に留まり、異常度自体はスパースではない。半教師あり学習の一種の未知データの異常度に直接スパース化を試みた例(スパース化異常度)は、知る限り無い。

通常の SMD は$D^2 = (y-\bar{x})'\mathrm{S}_x^{-1}(y-\bar{x}) = z'z$、student 化スコアベクトル$z = \mathrm{S}_x^{-1/2}(y-\bar{x})$と表せ、 zは線形方程式 $y-\bar{x} = \mathrm{S}_x^{1/2}z$ の解である。そこで、 zの線形方程式に LASSO を適用して求めたスパース化された student 化スコアベクトル $\hat{z}$を用いて、スパース化 SMD $(d^2 = \hat{z}'\hat{z})$ が得られる。LASSO 解法の CD (Coordinate Descent)法で $y-\bar{x} = \mathrm{S}_x^{1/2}z$ を解く。

スパース化$\hat{z}$を求める計算アルゴリズムは以下のとおり。

① 次元pの共分散行列S_xの固有値分解 $S_x = V_x'\mathrm{diag}(l_1, l_2, \cdots, l_p)V_x$ から平方根行列 $\mathrm{S}_x^{1/2} = \mathrm{diag}(\sqrt{l_1}, \sqrt{l_2}, \cdots, \sqrt{l_p})V_x$ を求める。

② CD 法による反復計算を$\hat{z}$が収束するまで実行する。

$$\hat{z}_{(j)} \leftarrow S_\rho\left(\frac{1}{p}\sum_{i=1}^{p}\mathrm{S}_x^{1/2}{}_{(i,j)}r_i^{(j)}\right)\Big/\frac{1}{p}\sum_{i=1}^{p}\left\{\mathrm{S}_x^{1/2}{}_{(i,j)}\right\}^2,$$

$$r_i^{(j)} = y_{(i)} - \bar{x}_{(i)} - \sum_{k\neq j}^{p}\mathrm{S}_x^{1/2}{}_{(i,j)}\hat{z}_{(k)},$$

$$S_\rho(x) = \mathrm{sign}(x)(|x|-\rho)_+, (y)_+ = \begin{cases} y\ (y \geq 0) \\ 0\ (y < 0) \end{cases}$$

yが学習データと同じ分布ならば、$\mathrm{S}_x^{1/2}$の固有値の範囲とyの大きさの大よその範囲は$\sqrt{l_1}$～$\sqrt{l_p}$となる。このアルゴリズムでは、$\mathrm{S}_x^{1/2}$とyの積がρ未満ならば soft-thresholding 関数$S_\rho(x)$により $\hat{z}_{(j)}$は 0 値になるから、大よそρ未満のS_xの固有値に当たる主成分スコア$\hat{z}_{(j)}$が強制的に 0 値になることがわかる。課題として、スパース化 SMD の数値誤差に対する安定性の理論的評価が必要である。

数値実験で、母固有値$\lambda_0 = 0$を 1 個含む母共分散行列をもつ$p = 7$次元の多変量正規分布に従うx, yを多数発生するモンテカルロシミュレーションを実行し、$\rho = 10^{-30} \cong l_0$として student 化スコアベクトル$z$の要素として$\lambda_0 = 0$に当たる標本主成分スコアの分布を得た。具体的には通常 SMD の主成分スコア$a(y) = (y-\bar{x})\cdot f_o/\sqrt{l_0}$ 、Ridge 正則化主成分スコア$b(y) = (y-\bar{x})\cdot f_o/\sqrt{l_0+\rho}$、スパース化 student 化スコアベクトル$\hat{z}$の要素$c(y) = \hat{z}_{(0)}$の分布を得た。正確には$a(y)$は 0 値のみを取るが、通常の SMD では数値誤差で 0 よりわずかに大きい標本固有値l_0を$a(y)$の分母にもつため、$a(y)$の分布は 0 値を中心に$\sqrt{l_0}$程度に広がり不安定化した。また、スパース化提案以前から不安定現象の除去を目的として用いられる Ridge 正則化したスコア$b(y)$は$a(y)$よりも狭いが 0 値を中心に分布していた。しかし、スパース化 SMD の$c(y) = \hat{z}_{(0)}$では$\rho \cong l_0$により数値誤差の影響を$\hat{z}$から除くことができ、正確に 0 値のみを取った。

また、ρより大きな固有値に当たる$\hat{z}_{(i)}$の分布は通常 SMD の対応する主成分スコアの分布に等しかった。図 1 の$\rho = 10^{-18}$の場合、固有値$\lambda_i \geq 10^{-10}$の主成分スコアのみ通常 SMD の主成分スコアと等しい分布となり、$\lambda_i = 10^{-15}$ではサンプルの微小値が 0 値へ集中し、$\lambda_i \leq 10^{-20}$では全サンプルが 0 値へ集中して、予想どおりの挙動を示した。

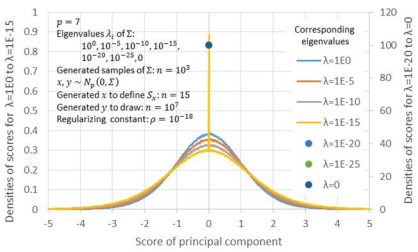


図 1　スパース化 SMD の主成分スコアの確率密度

提案したスパース化 SMD により、SMD の数値計算上の誤差に由来する不安定現象を除去可能となると期待される。具体的には正確には 0 値以外ありえない主成分スコアが SMD や Ridge 正則化 SMD では数値計算上の誤差により 0 値を中心に広がった分布をもつが、スパー

ス化 SMD では正確に 0 値のみをもつからである。課題として、SMD の数値計算上の誤差を除くために適切なρの値を事前に決める方法が必要である。

(2) 母固有値・母固有ベクトルの推定不要なスチューデント化主成分要素の近似式の提案 [2]

試験データyの SMD が大きい場合は、 yが学習データを近似する多変量正規分布から外れたものと判断できる。次はこの外れたyを構成する変数を定量的に検討する必要がある。p個の元の変数のままではなく、学習データからの共分散行列S_xの主成分要素$(y-\bar{x})\cdot f_i$（yの固有ベクトルf_i方向の大きさ）でyを表現して検討しても良い。

良く知られた方法の１つ目は、バイプロットである。バイプロットはyの散布図を、 p個のうち 2～3 個の主成分要素軸（多くは最大 1 位と 2 位の固有値に対応する主成分要素）で図示したものである。しかし、バイプロットでは 4 個以上の主成分要素軸により図示表現できず、また定量的な評価は難しい。

良く知られた方法の２つ目は、マハラノビス・タグチ・システム(MTS)に含まれる直交表解析である。2 水準の直交表に基づいて元の変数の選択を行ない、元の変数がもつマハラノビス距離の寄与を推定する。この直交表に基づく選択は直交表のもつ全組合わせで試行する必要があり非常に計算時間がかかる上、変数選択性能に欠点があるとの指摘もある。

良く知られた方法の３つ目は、主成分要素をもとにしたスチューデント化主成分要素（$z_i=(y-\bar{x})\cdot f_i/\sqrt{l_i}$ ）である。ちなみに SMD はスチューデント化主成分要素 z_i の 2 乗和で表せる。

$$SMD=\sum_{i=1}^{p}\left\{\frac{(y-\bar{x})\cdot f_i}{\sqrt{l_i}}\right\}^2=\sum_{i=1}^{p}z_i^2$$

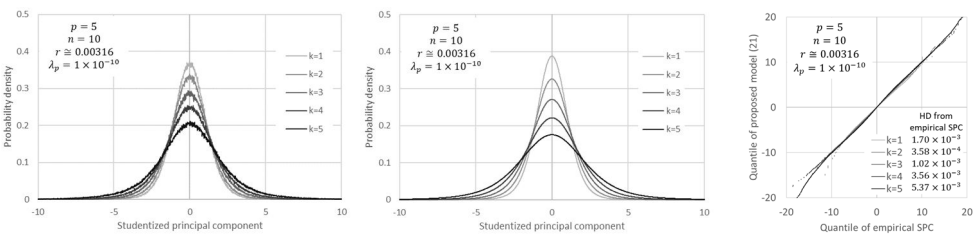
z_i の従う確率分布モデルとして、漸近モデル(正規分布による近似)や大標本モデル(t 分布による近似)が知られている。しかし、正規分布は標本数が非常に多い場合のみ有効であり、標本固有値の大きさの順位に応じて t 分布からのバイアス（降順の順位の高いl_iほど t 分布と比べて集中する傾向）が知られており、実用にはさらに正確なモデルが必要である。しかし、スチューデント化主成分要素 z_i は学習データのバラツキに起因してばらつく標本固有値l_iと固有ベクトルf_iを含むため、今まで正確なモデルは提案されてこなかった。

スチューデント化主成分要素z_iの含む標本固有値l_iと標本固有ベクトルf_iの展開式をz_iへ代入した式のモーメント母関数を検討し、 n個の学習データx_iと試験データyの従うp次元多変量正規分布の母共分散行列Σ_xの母固有値λ_iがお互いに十分に離れていて$(\lambda_1\gg\cdots\gg\lambda_p)$、さらに標本固有値の降順の順位$k$と比べて学習データ($p$次元)の標本数$n$が非常に大きければ、$z_k$が自由度$n-k$の t 分布$t(n-k)$に簡単な係数を乗じた次式に示す確率分布モデルで近似可能と示した。つまり母固有値λ_iと母固有ベクトルϕ_iの事前推定の必要がない。

$$z_k=\frac{f_k'\cdot(y-\bar{x})}{\sqrt{l_k}}\cong\frac{n-1}{n-k}\sqrt{\frac{n+k-2}{n-1}}\cdot t(n-k)$$

この式で$(n-1)/(n-k)$は標本固有値l_iのバラツキ、$\sqrt{(n+k-2)/(n-1)}$は標本固有ベクトルf_iのバラツキに由来する。これらの係数により、 t 分布からのバイアス（降順の順位の高いl_iほど t 分布と比べて集中する傾向）を正確な表現が可能になった。

モンテカルロシミュレーションにより 1000 万個発生させたz_iの確率密度分布(図 2-1(a))と上記の近似式モデルによる確率密度分布(図 2-1(b))を比較すると、いずれの主成分要素でも両者がよく合うことを確認できた。分布の裾の部分の差異を見るための両者の Q-Q プロット(図 2-1(c))では、どの主成分要素も直線$y=x$上によく乗り、シミュレーションで発生したz_iが上記近似モデルによく従うと確認できた。さらに両者の分布の近さを総合的に表すヘリンジャー距離(HD)もモンテカルロシミュレーションの誤差限界に近く、一致を示せた。また母固有値が$\lambda_1\gg\cdots\gg\lambda_p$の厳しい条件を満たさずに、隣接母固有値が近い(比 $r=\lambda_{i+1}/\lambda_i\cong 0.3\sim0.8$)場合でも、$n$と$p$が両方大きいと近似モデルによく従うことを確認できた。



(a) シミュレーションの結果　　(b) 提案近似モデルの結果　　(c) Q-Q プロット

図 2-1　シミュレーションと提案近似モデルによるz_iの確率密度分布 [2]

この近似モデルを利用し、近似モデルの累積分布関数から正規分布の逆累積分布関数を用いて正規分布の 2 乗和であるχ^2分布に変換するモデル(CSP)を考案した。

$$CSP = \sum_{k=1}^{p} \left\{ \Phi_{N(0.1)}^{-1} \left(\Phi_{t(n-k)} \left(\frac{f_k' \cdot (y - \bar{x})}{\sqrt{l_k}} \cdot \frac{n-k}{n-1} \sqrt{\frac{n-1}{n+k-2}} \right) \right) \right\}^2$$

図 2-2 に既存の補正方法と比較した結果を示す。既存の補正法として、pd は母固有値λ_iを事前推定する必要があるが、fLd は母固有値λ_iと母固有ベクトルϕ_iを事前推定する必要がない。図 2-2 左の確率密度グラフから、CSP のピークが補正目標である PMD のピークに最も近いことがわかる。さらに図 2-2 右の Q-Q プロットを比較すると pd や fLd のグラフよりも CSP のグラフが PMD へ近いが、同図 2-2 右のヘリンジャー距離を比較すると CSP よりも fLd の方が若干小さく PMD へ若干近い。この理由は、図 2-2 左では既存の補正法の pd や fLd で

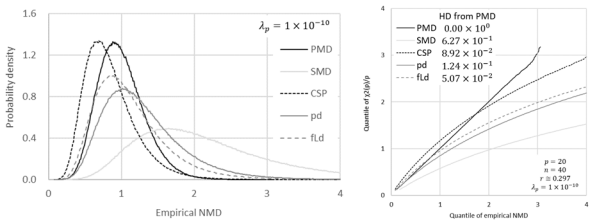


図 2-2　SMD($p = 20, n = 40$)から PMD への補正性能の比較 [2]

は CSP よりも確率密度グラフの右すそが右へ広がり、CSP のピークが PMD よりも小さい側にずれているためである。

＜引用文献＞

[1] Yasuyuki Kobayashi, “New Sparse Modeling of Sample Mahalanobis Distance”, Book of Abstracts Data Science, Statistics, and Visualization 2019 (DSSV2019), p.68, (2019), https://iasc-isi.org/dssv2019/programme/

[2] Yasuyuki Kobayashi, "New precise model of studentized principal components", Communications in Statistics - Theory and Methods, (2022), https://doi.org/10.1080/03610926.2022.2084110

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Kobayashi Yasuyuki	４．巻 Latest article
２．論文標題 New precise model of studentized principal components	５．発行年 2022年
３．雑誌名 Communications in Statistics - Theory and Methods	６．最初と最後の頁 1～18
掲載論文のDOI（デジタルオブジェクト識別子） 10.1080/03610926.2022.2084110	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計2件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 小林靖之
２．発表標題 [1C01-32-02] 標本マハラノビス距離の新しいスパースモデリング
３．学会等名 第15回日本統計学会春季集会
４．発表年 2020年～2021年

１．発表者名 Y. Kobayashi
２．発表標題 New Sparse Modeling of Sample Mahalanobis Distance
３．学会等名 DSSV (Data Science, Statistics & Visualization) 2019 (国際学会)
４．発表年 2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

帝京大学研究者総覧（ 小林靖之）
https://www3.med.teikyo-u.ac.jp/profile/ja.f8c1afded77c9900.html
researchmap（ 小林靖之）
https://researchmap.jp/ykoba1974

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

8．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関