Modal PCAの収束性とAngular Breakdown Pointの下界評価

日野　英逸　　モデリング研究系/統計的機械学習研究センター　教授

【概要】

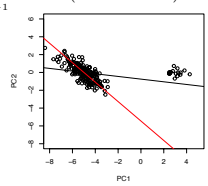
分布の分散ではなく最頻値に着目して外れ値に頑健な主成分分析（PCA）手法を開発した．手法の理論的性質として外れ値が混在しない場合には従来手法と同様の主成分を同定できることと従来手法と比較して外れ値の影響が限定的であることを示し，外れ値をどの程度許容できるかの計算可能な評価尺度を導出した．

【動機と意義】

・主成分分析（Principal Component Analysis, PCA）は次元削減，可視化，ノイズ除去を始めとして広く利用される多変量解析手法である．

・通常のPCA (classical PCA, **cPCA**)では標本分散を最大化する射影方向を抽出する：

$$\max_{v\in\mathcal{S}_{d-1}} \sum_{i=1}^{n}\left[v^\top x_i - \frac{1}{n}\sum_{j=1}^{n} v^\top x_j\right]^2 \Leftrightarrow \max_{v\in\mathcal{S}_{d-1}} v^\top X^\top \left(I - n^{-1}\mathbf{1}\mathbf{1}^\top\right) Xv$$



・分散は外れ値に大きく影響を受ける

・分散以外の尺度を用いた主成分分析手法がいくつか提案されている（Projection Pursuit）

・データが最もばらつかない方向は，「関心のない方向=minor component」であると考える

・Minor componentを，データを射影したときの**モード（最頻値）**の確率値が最大になる方向として定義

・外れ値の影響が少ない多変量解析手法は，データ取得コストが低い代わりに質が低いデータ解析のために必須．また，意図的に偽のデータを入れることで統計解析結果を歪める攻撃にも頑健な手法はセキュリティの観点からも重要

【アプローチ】

・データが一点に集まるほど大きくなり，かつ外れ値に頑健な量として，最頻値の確率密度値を採用．確率密度値が最大になる方向をminor component (MC) として推定し，その方向を「取り除く」ことで主成分のみを残す (**modal principal component analysis: mPCA**)

・MCの推定量

$$(\widehat{m}_k, \widehat{v}_k) = \arg\max_{m\in\mathbb{R},\ v\in\mathcal{S}^{d-1}} \frac{1}{N}\sum_{i=1}^{N}\phi_h\left(m - v^\top x_i\right),$$
$$\text{s.t.}\quad v^\top \widehat{v}_j = 0,\quad j = 1\ldots k-1.$$

$\phi_h(z)$ はカーネル関数で，$\phi_h(z) = \phi(z/h)/h$．以下では $\phi(z) = \exp\left(-z^2/2\right)/\sqrt{2\pi}$ とする

$\widehat{m}_k$ で MC_k 方向に射影された変数のモードの推定量を表す

【主結果1：Minor Componentの一様確率収束性】

・任意の射影方向におけるカーネルモード推定値は真の確率密度関数値に一様確率収束する．すなわち，幾つかの正則条件の下で次が成り立つ：

$$\sup_{(m,v)\in M\times\mathcal{S}^{d-1}}\left|\frac{1}{n}\sum_{i=1}^{n}\phi_h(m - v^\top X_i) - f_{v^\top X}(m)\right| = o_p(1)$$

ここで $(m_0, v_0) = \arg\sup_{m\in\mathbb{R},\ v\in\mathcal{S}^{d-1}} f_{v^\top X}(m).\ |m_0| < \infty \quad M = [-m_0, m_0]$

【主結果2：Minor Componentの一様確率収束レート】

・mPCAによる1st minor componentの一様確率収束レートは，モード推定に用いるカーネル関数のバンド幅の減少レートを $h_n = O(n^{-1/k}), k > 4$ とすると以下で与えられる．

$$\sup_{(m,\boldsymbol{v})\in M\times\mathcal{S}^{d-1}}\left|\frac{1}{n}\sum_{i=1}^{n}\phi_h(m - \boldsymbol{v}^\top \boldsymbol{X}_i) - f_{\boldsymbol{v}^\top \boldsymbol{X}}(m)\right| = O(n^{-1/k})$$

この結果は k 回微分可能な確率密度関数に対する既存のカーネルモード推定の収束レート $O((n/\log n)^{-k/(2k+1)})$ より遅い．モードと射影軸の同時一様確率収束を評価するために遅くなっている．

【主結果3：Finite-Sample Breakdown Point and its Lower Bound】

・あるa個のデータから求めた推定量を，適当なデータをb個加えることで任意に「悪く」できるような最小のbが大きければ大きいほど「頑健」な推定量

・PCA向けのbreakdown pointとしてangular breakdown pointを道入する：

$$\epsilon^*(\widehat{v}_k, Y_a) = \min_b\left\{\frac{b}{a+b}\middle|{}^{\exists}Y_b\subset\mathcal{Y},\ |Y_b| = b,\ \widehat{v}_k(Y_a\cup Y_b)^\top\widehat{v}_k(Y_a) = 0\right\}$$

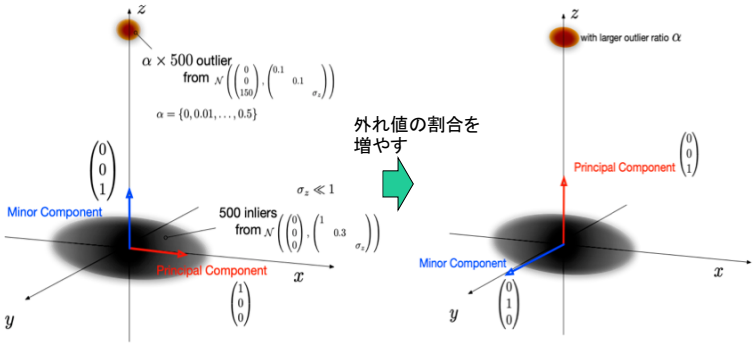
・Angular breakdown pointは主成分を「直交させる」ことができるデータの最小数を与える

・このbreakdown pointそのものは計算が難しいが，そのlower boundは観測データから計算可能：

$$\epsilon^*(\widehat{w}_1, Y_a) > \frac{b^*}{a+b^*},\quad \text{where}$$
$$\begin{cases} b^* = \lceil M_a(\widehat{w}_1(Y_a)) - M_a^*(\widehat{w}_1(Y_a))\rceil - 1, \\ M_a(\widehat{w}_1(Y_a)) = h\sqrt{2\pi}\sum_{i=1}^{a}\phi_h(\widehat{w}_1(Y_a)^\top x_i), \\ M_a^*(\widehat{w}_1(Y_a)) \\ \quad = \sup\left\{h\sqrt{2\pi}\sum_{i=1}^{a}\phi_h(w^\top x_i)\middle|\begin{matrix} w\in\mathcal{S}^{d-1}, \\ w^\top\widehat{w}_1(Y_a) = 0\end{matrix}\right\}\end{cases}$$

・手元のデータにどのくらい外れ値が入ってしまったら破綻するかを見積もることができる

・人工的に，minor componentベクトルの推定値が真の方向と直交しうるような外れ値を段階的に添加したときの，真の方向と推定された方向の角度（cosine）を評価する



・z軸方向の分散を適当に変化させることで，Lower Bound of Breakdown Point (LBBP)の値を変える．

・外れ値の混入率を0から0.5まで変化させ，サンプルを100回取り直して推定されたMinor Componentと真のMinor Componentのcosineを計算．

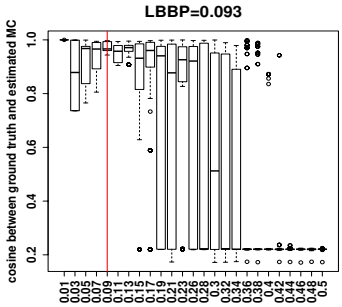
・LBBPの値は事前に50のinlierを用いて計算しておく

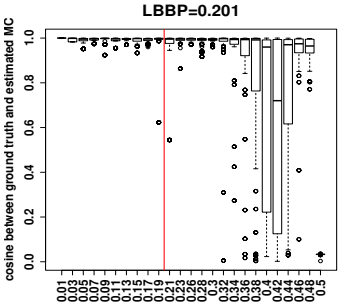
【結果】

・LBBPは実際に推定が破綻する混入率より小さい値を取っている．

→breakdown pointの保守的な評価

→実用上，少数のinlierのみを用いてLBBPを評価することで，どの程度の外れ値の混入が許容されるかを見積もることができる．





【課題】

・求解アルゴリズムの改善（計算効率向上）

←minor componentを取り除くことで

principal componentを残すという構造上，本質的に計算コストが高い

・Finite-sample breakdown pointの下界の活用

★本研究は筑波大学大学院システム情報工学研究科三戸圭史氏との共同研究です．

本発表に関する論文は，

・Neural Computation Vol.32(10), 2020

https://www.mitpressjournals.org/doi/abs/10.1162/neco_a_01308?journalCode=neco

・https://arxiv.org/abs/2008.03400

から取得可能です．また，実験に利用したソースコードは

https://github.com/KeishiS/ModalPCA

から取得可能です．