2019年6月5日　統計数理研究所　オープンハウス

モード推定に基づくロバスト主成分分析

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　准教授

【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．実験的にも，従来手法と同等以上のロバスト性を有することを確認した．

【動機と意義】

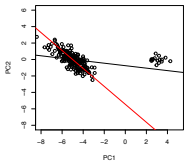
・主成分分析（Principal Component Analysis, PCA）は次元削減，可視化，ノイズ除去を始めとして広く利用される多変量解析手法である．

・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 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，$\phi_h(z)=\phi\left(z/h\right)/h$．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】

・mPCAによる1st minor componentは，cPCAによる1st minor component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 $(m_0,v_0)=\arg\sup_{m\in\mathbb{R},\ v\in\mathcal{S}^{d-1}} f_{v^\top X}(m)$．$|m_0|<\infty$，$M=[-m_0,m_0]$

←提案するmPCAは，cPCAの「妥当な」拡張になっている

【主結果2：Influence Function】

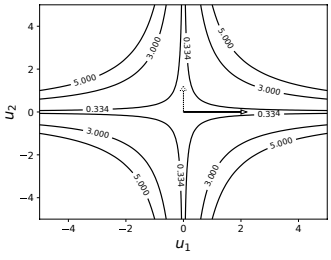
・推定量Tに対して一つの外れ値uがどの程度影響を及ぼすかを定量化(F, Δ_uはそれぞれデータが従う確率測度とディラック測度)

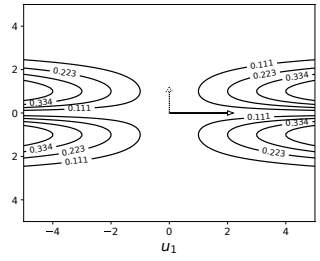
$$\mathrm{IF}(u;T,F)=\lim_{\epsilon\to 0}\frac{T\left((1-\epsilon)F+\epsilon\Delta_u\right)-T(F)}{\epsilon}$$

・mPCAによるk番目のminor componentの推定量 $\widehat{w}_1$ の影響関数：

$$\mathrm{IF}(u;\widehat{w}_k,F)=\left(A_kB_k-C_k\right)^{-1}\times\left[A_kd_k+\sum_{l=1}^{k-1}C_l\mathrm{IF}(u;\widehat{w}_l,F)\right],\ (k\geq 2),$$
$$\mathrm{IF}(u;\widehat{w}_1,F)=\left(A_1B_1-C_1\right)^{-1}A_1d_1$$

$$\begin{cases} A_k=I-\sum_{l=1}^{k}\widehat{w}_l(F)\widehat{w}_l(F)^\top, \\ B_k=\frac{1}{h^3}\int\left.\frac{d^2\phi(z)}{dz^2}\right|_{z=\frac{\widehat{w}_k(F)^\top x}{h}}xx^\top dF(x), \\ C_k=\widehat{w}_k(F)^\top\psi(\widehat{w}_k(F),F)I+\widehat{w}_k(F)\psi(\widehat{w}_k(F),F)^\top \\ d_k=\psi(\widehat{w}_k(F),F)-\frac{1}{h^2}\left.\frac{d\phi(z)}{dz}\right|_{z=\frac{\widehat{w}_k(F)^\top u}{h}}u, \\ \psi(w,F)=\frac{1}{h^2}\int\left.\frac{d\phi(z)}{dz}\right|_{z=\frac{w^\top x}{h}}xdF(x). \end{cases}$$





【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

・Finite-sample breakdown point

$$\epsilon^*(T,Y_a)=\min\left\{\frac{b}{a+b}\middle|\sup_{Y_b}|T(Y_a\cup Y_b)-T(X_a)|=\infty\right\}$$

・ロバスト統計の分野ではこの量が議論されることが多いが，PCAにおいてはTは単位ベクトルなので，この定義は使えない

・PCA向けのbreakdown pointの新しい定義：

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

・主成分を「直交させる」ことができるデータの最小数を考えている

・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

【最適化】

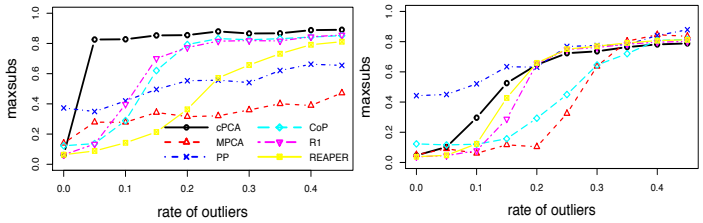
・適当な初期値から交互最適化を行う

・初期値の選択：GRIDアルゴリズム，モードの推定：Half-sample mode法，モードを固定したときの主軸の最適化：座標変換による多様体上の最適化

【実験】

・maxsub尺度で，外れ値なし（真値）の1st PCと推定量の角度を測って評価

・人工データ：20次元ガウス分布（図左：各次元で次元の逆数に反比例する分散を持つ），ラプラス分布（図右：次元の逆数に比例するスケール）から200点サンプリングして，一部を一様分布に従う「外れ値」で置き換える



・実データ：UCI machine learning repositoryから3種類のデータを取得

(inlier/outlier/dimension)

	mPCA	PP	CoP	R1	REAPER	cPCA
Parkinson(195/47/22)	0.00269	0.03355	0.01085	0.00718	0.0065	0.00645
Pendigits(9848/20/16)	0.06030	0.44240	0.72070	0.08630	0.1595	0.86870
Wilt(4562/257/5)	0.28500	0.15500	0.37200	0.34000	0.3130	0.36500

←他手法と比較してロバストであることを確認

【課題】

・求解アルゴリズムの改善（計算効率向上）

・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．