主成分分析

以下, 行列 $X = (x_{i,j}) \in \mathbb{R}^{N\times p}$ は中心化されていると仮定する．すなわち, 各 $j = 1, \cdots, p$ に対して

$$\sum_{i=1}^{N} x_{i,j} = 0$$

であるとする．$\|v\|_2^2 = 1$ のもとで,

$$\|Xv\|_2^2 = v^T X^T X v \tag{1}$$

を最大にする v を v_1，(**??**) 式を最大にして v_1 と直交する v を $v_2, \cdots$ というようにして正規直交系 $V := [v_1, \cdots, v_p]$ を求める操作を主成分分析という．

まず各 v_j が直交するという制約を除いて，$\|v\|_2^2 = 1$ のもとで $\|Xv\|_2^2$ を最大にする v_j について考えてみる．そのような v_j は

$$L(v_j, \mu_j) = \|Xv_j\|_2^2 - \mu_j(\|v_j\|_2^2 - 1)$$

を最大にするので，上式を v_j で微分して 0 とおいた等式

$$X^T X v_j - \mu_j v_j = 0$$

を満足する．上式を，X の標本共分散行列 $\Sigma := \frac{1}{N} X^T X$ および $\lambda := \frac{\mu_j}{N}$ を用いて書き換えると

$$\Sigma v_j = \lambda_j v_j$$

と書くことができる．これより求める v_j は Σ の固有ベクトルで，λ_j は v_j が属する固有値になることがわかる．もし λ_j の中で重複度が 2 以上のものだある場合には，それらの固有ベクトルは直交するように選んでくる．Σ の固有値が全てことなる場合には，$v_1, \cdots, v_p$ は自動的に直交することがいえる．

実際には $v_1, \cdots, v_p$ を全て用いることはなく，最初の m 個のみを用いることになる．そして X の各行を $V_m := [v_1, \cdots, v_m]$ に射影した $Z := XV_m$ を得る．すなわち p 次元の情報を m 個の主成分 $v_1, \cdots, v_m$ の空間に射影して，m 次元の Z で p 次元の X を見ることになる．そのような次元の圧縮のための線形写像が，主成分分析である．

主成分分析のスパースなアプローチにはいくつかあるので，それらを考察していく．まず非ゼロ要素の個数を制限する手法について，この場合は t を整数として，$\|v\|_0 \leq t, \|v\|_2 = 1$ のもとで

$$v^T X^T X v - \lambda \|v\|_0$$

を最大化するような定式化になる．しかしこの場合，目的関数が凸にはならない．

また，$\|v\|_1 \leq t (t > 0)$ の制約を持たせて，$\|v\|_2 = 1$ のもとで

$$v^T X^T X v - \lambda \|v\|_1 \tag{2}$$

の最大化を図ろうとしても，目的関数は凸にならない.

そこで $u \in \mathbb{R}^N$ として，$\|u\|_2 = \|v\|_2 = 1$ のもとで

$$u^T X v - \lambda \|v\|_1 \tag{3}$$

の最大化を図る定式化，SCoTLASS[*1]が提案された．(**??**) 式で得られる最適な v は，(**??**) の最適解になっている．実際

$$L := -u^T X v + \lambda \|v\|_1 + \frac{\mu}{2}(u^T u - 1) + \frac{\delta}{2}(v^T v - 1) \tag{4}$$

を u で偏微分して 0 とおくと

$$\frac{\partial L}{\partial u} = Xv + \mu u = 0$$

となり，$\|u\|_2^2 = 1$ であることから $u = \dfrac{Xv}{\|Xv\|_2}$ となる．これを (**??**) 式に代入することで

$$-\|Xv\|_2 + \lambda \|v\|_1 + \frac{\delta}{2}(v^T v - 1) \tag{5}$$

となる.

[*1] Simplified Component Technique - LASSO