Advances in Applied Mathematics 应用数学进展, 2017, 6(9), 1174-1181
Published Online December 2017 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2017.69142

Hans汉斯

Two-Stage Method of Sparse Principal Component Analysis

Xin Yang

The School of Mathematics and Systems, Beihang University, Beijing
Email: msyangxin@buaa.edu.cn

Received: Dec. 1st, 2017; accepted: Dec. 19th, 2017; published: Dec. 26th, 2017

Abstract

In this paper, we propose a sparse principal component based on two-stage method, that is, we first get principal component, and then add the ℓ_1 regular term of the loadings. Coordinate descent method is used to solve the model. The model is easy to understand. In addition, this paper proposes a heuristic algorithm which can determine the penalty parameters in the model. By selecting the appropriate penalty parameters, the sparse principal component explained variance and sparsity can be optimized at the same time.

Keywords

Dimension Reduction, Sparse Principal Component Analysis, Least Absolute Shrinkage and Selection Operator, Coordinate Descent Method

稀疏主成分分析的两阶段法

杨　欣

北京航空航天大学数学与系统科学学院，北京
Email: msyangxin@buaa.edu.cn

收稿日期：2017年12月1日；录用日期：2017年12月19日；发布日期：2017年12月26日

摘　要

本文提出稀疏主成分分析的两阶段法，即先求解主成分，然后添加ℓ_1正则化项得到稀疏载荷，并利用坐标下降法求解模型。方法简单易操作。另外，本文还提出了一种可以确定两阶段模型中惩罚参数的算法，通过选取合适的惩罚参数，可以使稀疏主成分方差和主成分相关性等性能指标取得折衷。

文章引用: 杨欣. 稀疏主成分分析的两阶段法[J]. 应用数学进展, 2017, 6(9): 1174-1181.
DOI: 10.12677/aam.2017.69142

关键词

降维，稀疏主成分分析，最小迭代收缩阈值和选择算子，坐标下降法

1. 引言

多数情况下多指标问题中不同的指标之间有一定的相关性。指标较多加上指标之间的相关性，势必增加了问题分析的复杂性。主成分分析(Principal Component Analysis, PCA)就是一种通过降维技术把多个变量简化为少数几个互不相关的主成分的统计分析方法，主成分是原始变量的线性组合，同时能最大程度上保留原始数据的信息(信息含量用方差表示)。

由于主成分是原始变量的线性组合，并且主成分载荷元素通常非零[1]，因此很难对单个主成分做出解释。稀疏主成分分析(Sparse Principal Component Analysis, SPCA)就是为了解决这一缺陷而提出的方法，通过增加主成分载荷中零元素个数，使得主成分可以用最少且最有代表性的变量的线性组合来表示。

为了得到稀疏主成分，研究者们做了很多尝试。Cadima 等[1]利用硬阈值法，将主成分载荷中绝对值小于给定阈值的元素截断为 0，提高了主成分的可解释性，但是这种方法给出的主成分容易识别错误的原始变量。Hausman [2]将载荷的取值固定在离散集中，例如{-1,0,1}，但是该方法得到的载荷稀疏性并不理想，并且稀疏主成分的方差大大减小。2003 年 I.T.Jolliffe [3]受 LASSO [4]的启发，直接将 ℓ_1 范数约束引入到主成分模型当中，得到的 SCoTLASS 算法，该算法是第一个基于优化的稀疏 PCA 算法，但是该算法得到的主成分方差并不理想。在此之后又陆续出现了利用 ℓ_0 范数作为约束和 ℓ_0 范数、ℓ_1 范数作为惩罚项的稀疏 PCA 模型及算法。

本文提出稀疏主成分分析的两阶段法，第一阶段得到主成分，第二阶段为了得到稀疏主成分，利用 LASSO [4]问题可以使得回归系数自动缩减为 0 的特点，将稀疏主成分求解表述为 LASSO 问题，并用坐标下降法 CD [5] [6]求解该模型，方法简单易操作。此外，本文给出了一种有效确定惩罚参数的方法，利用该算法可以得到合适的惩罚参数，使得主成分各性能指标取得折衷。

下面列出本文中常用的符号术语。

符号 $\mathbb{R}^{n\times p}$ 表示 $n\times p$ 维实数空间，$\boldsymbol{S}^{n\times p}$ 表示 $n\times p$ 阶实对称矩阵。$\operatorname{sign}(\cdot)$ 表示符号算子。加粗小写字母表示一个向量，下标表示其分量，向量 $\boldsymbol{v}$ 的第 j 个分量用 v_j 表示，同样地，加粗大写字母表示一个矩阵，矩阵 $\boldsymbol{V}$ 的第 j 个列向量用 $\boldsymbol{v}_j$ 表示。符号 $\|\boldsymbol{v}\|_1=\sum_j|v_j|$ 表示向量 $\boldsymbol{v}$ 的 ℓ_1 范数。符号 $v_+=\max(v,0)$。符号 $\boldsymbol{S}_\lambda(v)=\operatorname{sign}(v)(|v|-\lambda)_+$ 表示 v 的软阈值算子。符号 $\|\boldsymbol{X}\|_F=\operatorname{tr}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})$，其中 $\operatorname{tr}(\bullet)$ 表示矩阵的迹。$\boldsymbol{I}_k$ 表示 k 阶单位矩阵。数据矩阵用 $\boldsymbol{X}\in\mathbb{R}^{n\times p}$ 表示，已中心化(列均值为 0)，样本协方差矩阵 $\Sigma=\dfrac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}}{n-1}\in\boldsymbol{S}^{p\times p}$，其中 p 表示变量个数，n 表示样本数量。

2. 稀疏主成分分析的两阶段法

本节简单说明与两阶段法密切相关的 SPCA 模型及算法与 LASSO 模型。

2.1. LASSO 问题与 SPCA 算法

回顾 LASSO 问题[4]

$$\beta^* = \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{2}\|\boldsymbol{X}\beta - \boldsymbol{y}\|_2^2 + \lambda\|\beta\|_1,$$

其中 $\boldsymbol{X}\in\mathbb{R}^{n\times p}$ 表示数据矩阵，而 $\boldsymbol{y}$ 是相应的观测值，λ 是控制 β 稀疏性的非负参数，通过选取合适的参数 λ 可以得到满足稀疏性要求的向量解 β。

Zou 等[7] 2012 年提出 SPCA 模型，SPCA 模型先将 PCA 表述为回归优化问题，然后添加关于回归系数的 ℓ_1 正则项，即

$$\begin{aligned} &\underset{A,B}{\text{minimize}} \quad \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^{\mathrm{T}}\|_{\mathrm{F}}^2 + \lambda\sum_{j=1}^{k}\|\beta_j\|^2 + \sum_{j=1}^{k}\lambda_{1,j}\|\beta_j\|_1 \\ &\text{subject to} \quad \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} = \boldsymbol{I}_k. \end{aligned} \tag{1}$$

其中正参数 $\lambda_{1,j}$ 控制载荷 β_j 的稀疏性。对于 $n>p$ 的数据集，要求 $\|\beta_j\|^2$ 前的参数 $\lambda>0$；对于 $n\le p$ 的数据集，取 $\lambda=0$。

SPCA 利用块坐标下降法将(1)的变量分成 $\boldsymbol{B}$ 和 $\boldsymbol{A}$ 两个坐标块，固定其中一个坐标块，求解关于另一个坐标块的子问题，交替求解关于两个变量的子问题直至满足终止条件。SPCA 模型将 PCA 与回归分析建立联系，并且对于不同的数据类型($n>p$ 或 $n\le p$)都具有较低的计算复杂度[7]。

SPCA 算法初值 $\boldsymbol{A}_0=\boldsymbol{B}_0$ 均取前 k 个主成分载荷 $\bar{\boldsymbol{v}}_1,\cdots,\bar{\boldsymbol{v}}_k$，固定 $\boldsymbol{A}$ 求解问题(1)等价于求解

$$\underset{B}{\text{minimize}} \quad \|\boldsymbol{X}\boldsymbol{A} - \boldsymbol{X}\boldsymbol{B}\|_{\mathrm{F}}^2 + \lambda\sum_{j=1}^{k}\|\beta_j\|^2 + \sum_{j=1}^{k}\lambda_{1,j}\|\beta_j\|_1, \tag{2}$$

令 $\boldsymbol{y}_j = \boldsymbol{X}\bar{\boldsymbol{v}}_j, j=1,\cdots,k$，求解(2)等价于求解 k 个独立的弹性网问题

$$\beta_j^* = \arg\min_{\beta_j}\|\boldsymbol{y}_j - \boldsymbol{X}\beta_j\|^2 + \lambda\|\beta_j\|^2 + \lambda_{1,j}\|\beta_j\|_1,\ j=1,\cdots,k. \tag{3}$$

2.2. 两阶段法的第一阶段与第二阶段

稀疏 PCA 的两阶段法第一阶段需要求得主成分，通常有两种方法可以得到主成分。

(1) 样本协方差阵特征值分解

由于实际数据中总体协方差阵未必已知，因此用样本协方差阵代替总体协方差阵。对样本协方差阵 Σ 做特征值分解，并将特征值降序排列，那么前 k 大特征值对应的特征向量即为前 k 个主成分载荷 $\bar{\boldsymbol{v}}_1,\cdots,\bar{\boldsymbol{v}}_k$，对应的主成分表示为 $\boldsymbol{y}_i = \boldsymbol{X}\bar{\boldsymbol{v}}_i, i=1,\cdots,k$。

(2) 数据矩阵奇异值分解

假设数据矩阵 $\boldsymbol{X}$ 有 SVD 分解为 $\boldsymbol{X}=\bar{\boldsymbol{U}}\bar{\boldsymbol{D}}\bar{\boldsymbol{V}}^{\mathrm{T}}$，其中 $\bar{\boldsymbol{U}}=(\bar{\boldsymbol{u}}_1,\cdots,\bar{\boldsymbol{u}}_r)\in\mathbb{R}^{n\times r}$，$\bar{\boldsymbol{V}}=(\bar{\boldsymbol{v}}_1,\cdots,\bar{\boldsymbol{v}}_r)\in\mathbb{R}^{p\times r}$，$\bar{\boldsymbol{D}}$ 是对角线元素为 $\sigma_1,\cdots,\sigma_r$ 的对角矩阵，$r=\mathrm{Rank}(\boldsymbol{X})$，$\bar{\boldsymbol{U}},\bar{\boldsymbol{V}}$ 的列均标准正交。令 $\bar{\boldsymbol{D}}$ 的对角线元素降序排列，在等式 $\boldsymbol{X}=\bar{\boldsymbol{U}}\bar{\boldsymbol{D}}\bar{\boldsymbol{V}}^{\mathrm{T}}$ 两边同时乘以矩阵 $\bar{\boldsymbol{V}}$，得到 $\boldsymbol{X}\bar{\boldsymbol{V}}=\bar{\boldsymbol{U}}\bar{\boldsymbol{D}}$，即矩阵 $\bar{\boldsymbol{U}}\bar{\boldsymbol{D}}$ 的列向量为主成分，$\bar{\boldsymbol{V}}$ 的列向量为对应的主成分载荷。不难看出，由于 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}=\bar{\boldsymbol{V}}\bar{\boldsymbol{D}}^2\bar{\boldsymbol{V}}^{\mathrm{T}}$，则经过数据矩阵 $\boldsymbol{X}$ 的奇异值分解得到的 $\bar{\boldsymbol{V}}$ 的列向量也是样本协方差矩阵 Σ 的特征向量。

将第一阶段得到的前 k 个主成分表示为 $\boldsymbol{y}_i = \boldsymbol{X}\bar{\boldsymbol{v}}_i, i=1,\ldots,k$。为了得到稀疏主成分，利用 LASSO [4] 问题可以使得回归系数自动缩减为 0 的特点，用稀疏主成分 $\boldsymbol{X}\boldsymbol{v}_i$ 拟合主成分 $\boldsymbol{X}\bar{\boldsymbol{v}}_i$，得到 k 个 LASSO 问题，即

$$\underset{\boldsymbol{v}\in\mathbb{R}^p}{\text{minimize}}\ \frac{1}{2}\left\|\boldsymbol{X}\boldsymbol{v}-\boldsymbol{y}_i\right\|_2^2+\lambda\left\|\boldsymbol{v}\right\|_1, i=1,\cdots,k,$$

可以看出这 k 个 LASSO 问题相互独立，因此只需要给出第一个 LASSO 问题的算法，其余 $k-1$ 问题以此类推。不失一般性，省略 $\boldsymbol{y}_i$ 的下标，用 $\boldsymbol{y}$ 表示。

$$\underset{\boldsymbol{v}\in\mathbb{R}^p}{\text{minimize}}\ \frac{1}{2}\left\|\boldsymbol{X}\boldsymbol{v}-\boldsymbol{y}\right\|_2^2+\lambda\left\|\boldsymbol{v}\right\|_1. \tag{4}$$

分别求解 k 次问题(4)，可以得到前 k 个稀疏载荷 $\boldsymbol{V}=\left(\boldsymbol{v}_1,\cdots,\boldsymbol{v}_k\right)$。

可以看到，问题(4)和(3)相差 $\lambda\left\|\boldsymbol{v}_j\right\|^2$ 这一项，但是在求解 $n\geq p$ 稀疏 PCA 问题时，通常令 $\lambda=0$。因此(4)是 SPCA 准则(1)第一次迭代时所求解的 k 个子问题。由于 SPCA 算法同时需要以主成分载荷作为初值，因此这部分计算量与两阶段法第一阶段计算量相同，其次两阶段法的第二阶段是分别求解 k 个子问题，而 SPCA 需要交替求解关于变量 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的子问题，因此计算量要大于两阶段法，第 3 节的数值结果也证实了这一点。

问题(4)是一个标准 LASSO [4]问题，从而使用经典的求解(3)的算法来求解。文献[7]中利用内点法进行求解，但针对基因数据等大规模数据，内点法计算时间过大，效率低下。由于坐标下降法[6]能够充分利用高维数据的稀疏特性，已经成为处理大规模稀疏数据的首选算法。因此，本文选择坐标下降法求解问题(4)。

2.3. 基于 LASSO 的坐标下降法

坐标下降法[5]在当前点处沿一个坐标方向进行一维搜索，同时固定其他坐标方向，求解目标函数的局部极小值。在给出坐标下降法求解问题(4)之前，先给出命题 1。

命题 1：对于任意的 $\alpha\in\mathbb{R},\lambda>0$，$\frac{\boldsymbol{S}_\lambda(\alpha)}{b}$ 表示函数 $\frac{1}{2}bx^2-\alpha x+\lambda\left|x\right|$ 的最小值点，其中 $\boldsymbol{S}_\lambda(\alpha)$ 表示软阈值算子，即

$$\boldsymbol{S}_\lambda(\alpha)=\begin{cases}\alpha-\lambda, & \alpha>\lambda\\ 0, & \left|\alpha\right|\leq\lambda\\ \alpha+\lambda, & \alpha<-\lambda\end{cases}$$

将坐标下降法运用到两阶段法模型(4)。针对 $\boldsymbol{v}$ 的第 i 个分量 v_i 求解问题(4)，同时固定 $v_j\left(j\neq i\right)$ 的值不变，即相当于求解

$$v_i^*=\arg\min_v\left\{\sum_{j=1}^n x_{ji}^2v_i^2-\left(\sum_{j=1}^n x_{ji}r_j^{(i)}\right)v_i+\lambda\left|v_i\right|\right\},$$

其中 $r_j^{(i)}=y_j-\sum_{k\neq i}x_{jk}v_k$。根据命题 1，得出该子问题的解为

$$v_i=\frac{\boldsymbol{S}_\lambda\left(\sum_{j=1}^n x_{ji}r_j^{(i)}\right)}{\sum_{j=1}^n x_{ji}^2}, \tag{5}$$

因此坐标下降法的第 i 步迭代即为(6)。文献[5]中称这种更新方法为平凡更新(naïve updating)，并给出了该方法的计算复杂度为 $O(np)$。

对于 $n\gg p$ 的数据集，显然平凡更新计算量很大，因此[5]中还给出了另一种协方差更新(covariance updating)方法，即令

$$v_i = \frac{\boldsymbol{S}_\lambda\left(\langle \boldsymbol{x}_i, \boldsymbol{y}\rangle - \sum_{m:|v_m|>0} \langle \boldsymbol{x}_i, \boldsymbol{x}_m\rangle v_m + v_i\right)}{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i}, \tag{6}$$

文献[5]中给出该方法的计算度仅为 $O(pm), m \le p \ll n$， 显然对于 $n \gg p$ 的数据集，后一种迭代方法计算复杂度更小。算法终止条件选取迭代次数小于最大迭代次数，或者目标函数值的更新率低于给定阈值，将求解(4)的算法称为 TSPCA。

2.4. 参数选择算法

LASSO 问题中的非负惩罚参数通常使用交叉验证[3]进行确定，但是这种方法高度依赖于数据矩阵，在已知样本协方差矩阵的数据集中，交叉验证的方法并不适用，因此本文提出一种有效的参数选择方法。稀疏度用 s 表示，下面给出参数选择算法 1。

算法 1 参数选择算法

1.for $s = 0,1,\cdots,p-1$ do
2.利用 TSPCA, 得到对应 s 的稀疏载荷 $\boldsymbol{v}$。
3.计算稀疏主成分的可解释方差百分比；
3. end for
4.画出稀疏主成分方差与稀疏度的函数图，选取方差变化最剧烈的一点(5%~10%)对应的稀疏度作为最佳稀疏度，对应的参数值即为最合适的参数值。

为了说明算法 1 的执行过程，以 Pitprop 数据为例。Pitprop 数据包含 180 个样本，13 个变量，$n = 180, p = 13$，实验提取 Pitprop 数据前 6 个稀疏主成分，即令 $k = 6$。图 1 是利用算法 1 得到 Pitprop 数据前 6 个主成分稀疏度(sparsity)与可解释方差百分比(PEV)函数图(见图 1)。

从图 1 分析可知，Pitprop 数据前 6 个稀疏主成分最佳稀疏度分别是 8，11，9，9，12，11，相应的参数分别 0.20，0.21，0.4，0.3，0.29，0.56，利用算法 1 得到了最佳参数。

3. 数值结果

本节将通过三组实验数据验证两阶段法的有效性和可行性，为了观察两阶段法(Two-stage method for Sparse PCA, TSPCA)得到的主成分不相关性和载荷非正交性，将 TSPCA 数值结果与 SPCA [7]和 ALSPCA [8]进行对比，为了观察 TSPCA 得到的主成分可解释方差，将 TSPCA 算法结果与 GPower [9]进行对比，其中 SPCA 算法代码由本文作者编写，GPower 和 ALSPCA 算法代码则分别下载自作者主页[1]。实验数据集包括 Pitprop 数据集，结肠癌基因数据集，20 新闻组数据集。所有数值实验均在处理器为因特尔四核 2.60Hz 的计算机上运行，所需代码利用 Matlab2012b 软件编写。

实验性能指标选择稀疏度，载荷非正交性[8]，稀疏主成分相关性[8]及 PEV [7]、计算时间，其中计算时间用来测试 TSPCA 算法处理大规模稀疏 PCA 问题的效率，均以秒为单位；稀疏度指的是稀疏载荷中零元素个数(当元素绝对值小于 0.001 时，即看作是零)。最大迭代次数控制为 1000，更新率阈值取作 1e-4，初始点 $\boldsymbol{v}_0$ 取主成分载荷，惩罚参数 λ 均利用算法 1 进行选取。

3.1. Pitprop 数据

本节利用 Jeffers [10] 1967 年引进的 Pitprop 数据测试 TSPCA 算法表现，它是主成分存在解释困难最为经典的数据集。Pitprop 数据包含 180 个样本，13 个变量，即 $n = 180, p = 13$，且前 6 个主成分载荷中没

[1]GPower 算法代码下载自 http://www.inma.ucl.ac.be/~richtarik. ALSPCA 算法代码下载自 http://www.math.sfu.ca/~zhaosong.

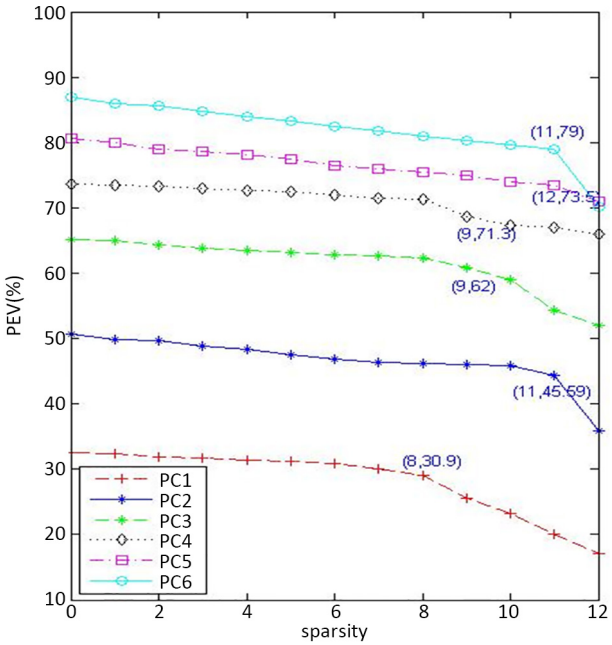

Figure 1. Six PCs of pitprop data: trade-off curve
图 1. Pitprop 数据：前 6 个稀疏主成分稀疏度随可解释方差百分比变化图

有零元素，因此本实验只提取前 6 个稀疏主成分。基于该数据，运行各稀疏 PCA 算法。根据算法 1，TSPCA 惩罚参数 λ 如 2.4 节所述。表 1 是 Pitprop 数据下 PCA 与各稀疏 PCA 算法的性能指标。

从表 1 比较发现，当载荷稀疏度相同，TSPCA 算法的 PEV 均高于除 GPower 以外的其他算法；在主成分相关性和载荷正交性方面，除了 ALSPCA 算法和 SPCA 算法之外，TSPCA 算法得到的稀疏主成分相关性最小；从计算时间的角度，TSPCA 所用时间均比其他算法更短，综合来看，TSPCA 算法有一定的竞争力。

以下两组实验是观察 TSPCA 解决大规模稀疏 PCA 问题的表现。

3.2. 结肠癌基因数据

结肠癌基因表达数据集[11]是大规模高维低样本数据，包含 62 个样本(22 个正常组织样本和 40 个癌变组织样本)和 2000 个基因表达变量，即 $n=62, p=2000$。由于前 3 个主成分的 PEV 大于 60%，因此本节抽取结肠癌基因数据前 3 个主成分。先将数据矩阵中心化。基于该数据，运行各稀疏 PCA 算法。利用算法 2，TSPCA 算法前 3 个惩罚参数均取 $\lambda=0.5, 0.4, 0.5$。表 2 是结肠癌基因数据下 PCA 与各稀疏 PCA 算法的性能指标。

从表 2 可以看出，TSPCA 算法所得的载荷在保持高稀疏度的情况下，虽然主成分 PEV 低于其他算法，但是 TSPCA 算法需要更少的计算时间，并且在非正交性和相关性方面表现也优于除 ALSPCA 以外的其他算法，验证了本文两种算法在解决高维稀疏 PCA 问题的简单有效。

3.3. 20 新闻组数据

本节所用的 20 新闻组数据，记录了 100 个单词出现在 16,242 篇新闻报导的频次，即 $n=16242, p=100$。所有的新闻报导均从全球最大的电子布告栏系统 Usenet 上取得。该数据下载自 http://blog.csdn.net/imstudying/article/details/77876159#OLE_LINK2。

基于该数据，运行各稀疏 PCA 算法，并抽取前两个稀疏主成分载荷的数据结果进行对比。对于

Table 1. Six sparse PCs of pitprop data: test index of sparse PCA algorithms
表 1. Pitprop 数据前 6 个稀疏主成分：各稀疏 PCA 算法的性能指标

算法	稀疏度	PEV(%)	非正交性	相关性
PCA	0	86.90	0	0
TSPCA	60	79	1.00	0.39
ALSPCA	60	79.42	0	0.30
GPower	60	79.74	17.88	0.91
SPCA	60	75.82	0.86	0.40

Table 2. Three sparse PCs of colon cancer data: Test index of sparse PCA algorithms
表 2. 结肠癌基因数据前 3 个稀疏主成分：各稀疏 PCA 算法的性能指标

算法	稀疏度	PEV(%)	非正交性	相关性	计算时间(s)
PCA	0	58.35	0	0.09	-
TSPCA	5377	49.35	5.36	0.40	0.3415
ALSPCA	5252	48.94	0.03	0	1.0000
GPower	5218	49.35	23.19	0.69	6.8507
SPCA	5370	48.88	22.88	0.49	5.4494

Table 3. Two sparse PCs of 20 news group data: test index of sparse PCA algorithms
表 3. 结肠癌基因数据前 3 个稀疏主成分：各稀疏 PCA 算法的性能指标

算法	稀疏度	PEV(%)	非正交性	相关性	计算时间(s)
PCA	0	10.69	0	0	--
TSPCA	160	8.11	1.15	0.20	0.6704
ALSPCA	160	8.48	0	0.10	35.6041
GPower	160	8.26	20.76	0.49	27.4600
SPCA	160	8.39	0.17	0.15	2.9733

TSPCA 算法，利用算法 2 取惩罚参数 $\lambda = 0.104, 0.187$。表 3 为 20 新闻组数据下 PCA 和各稀疏 PCA 算法的性能指标。

从表 3 可以看出，TSPCA 算法所得的载荷 PEV 虽然略低于其他算法，但是在非正交性和正交性的表现要优于 GPower 算法；就计算时间而言，TSPCA 算法用时最短，进一步验证了本文算法在解决大规模稀疏 PCA 问题方面的有效性。

4. 结论

本文首次提出稀疏 PCA 的两阶段法，并利用坐标下降法求解该模型，其次还提出一种可以选取最佳惩罚参数的方法。模型简单易懂，算法易于实现，同时两阶段法每次迭代的计算复杂度关于样本个数 n 和变量维数 p 都是线性的，因此可以有效求解大规模稀疏 PCA 问题，同时惩罚参数选取算法可以有效选取惩罚参数，使得载荷稀疏度和稀疏主成分可解释方差等指标取得折衷。

致　谢

衷心感谢指导老师和各位评阅人的建议！

参考文献 (References)

[1] Cadima, J. and Jollife, I.T. (1995) Loading and Correlations in the Interpretation of Principal Components. *Journal of Applied Statistics*, **22**, 203-214. https://doi.org/10.1080/757584614

[2] Hausman, R. (1982) Constrainted Multivariate Analysis in Optimization in Statistics. North Holland, Amsterdam, 137-151.

[3] Jolliffe, I.T., Trendafilov, N.T. and Uddin, M. (2003) A Modified Principal Component Technique Based on the LASSO. *Journal of Computational and Graphical Statistics*, **12**, 531-547. https://doi.org/10.1198/1061860032148

[4] Tibshirani, R. (1996) Regression Shrinkage and Selection via Lasso. *Journal of the Royal Statistical Society*, *Series B* (*Methodological*), **58**, 267-268.

[5] Hastie, T., Tibshirani, R. and Wainwright, M. (2016) Statistical Learning with Sparsity. A Chapman & Hall Book.

[6] Hsieh, C.J., Chang, K.W., Lin, C.J., *et al*. (2008) A Dual Coordinate Descent Method for Large-Scale Linear SVM. *Proceedings of the* 25*th International Conf on Machine Learning*, ACM, New York, 408-415.

[7] Zou, H., Hastie, T. and Tibshirani, R. (2012) Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, **15**, 265-286. https://doi.org/10.1198/106186006X113430

[8] Lu, Z. and Zhang, Y. (2012) An Augmented Lagrangian Approach for Sparse Principal Component Analysis. *Mathematical Programming*, **135**, 149-193. https://doi.org/10.1007/s10107-011-0452-4

[9] Journee, M., Nesterov, Y. and Richtarik, P. (2010) Generalized Power Method for Sparse Principal Component Analysis. *The Journal of Machine Learning Research*, **11**, 517-553.

[10] Jeffers, J.N.R. (1967) Two Case Studies in the Application of Principal component Analysis. *Applied Statistics*, **16**, 225-236. https://doi.org/10.2307/2985919

[11] Alon, U., Barkai, N., Notterman, D.A., *et al*. (1999) Broad Patterns of Gene Expression Revealed by Clustering of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 6745-6750. https://doi.org/10.1073/pnas.96.12.6745