一种稀疏可控的主成分分析方法

谭亚芳 刘 娟 王才华 蒋万伟
(武汉大学计算机学院 武汉 430072)

摘 要 主成分分析(Principal Component Analysis,PCA)是一种用线性变换选出少数重要变量(降维)的多元统计分析方法。虽然传统PCA被广泛应用于科学研究与工程领域中,但是其结果有时很难解释。因此,一些研究人员引入稀疏约束项(lasso、fused lasso以及adaptive lasso等),以得到可解释的结果。由于传统稀疏项的稀疏度不容易控制,为此引入一种新的约束项,即稀疏可控惩罚项(Sparse Controllable penalty,SCP),来控制主成分的稀疏程度。与传统的约束项相比,SCP具有长度不敏感、维度不敏感和约束项的取值范围在0到1之间的优点。这些优点极大地降低了调节稀疏度的难度。实验表明,稀疏可控主成分分析(Sparse Controllable Principal component Analysis,SCPCA)是高效的。

关键词 主成分分析,稀疏约束项,稀疏可控主成分分析
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.01.045

Sparse Controllable Principal Component Analysis Method

TAN Ya-fang LIU Juan WANG Cai-hua JIANG Wan-wei
(School of Computer,Wuhan University,Wuhan 430072,China)

Abstract Principal component analysis (PCA) is a multivariate statistical analysis method which chooses a few important variables (dimension reduction) by linear transformation. PCA is widely used in scientific researches and engineering,however,the results can sometimes be difficult to interpret. Therefore,some researchers introduced sparse penalties (lasso,fused lasso and adaptive lasso etc.) to obtain interpretable results. Since the traditional sparse penalty is not easy to control,we presented a novel penalty,namely sparse controllable penalty (SCP),to control the sparsity of principal components. Compared with the traditional penalties,SCP is scale insensitive,dimension insensitive and bounded between 0 and 1. It is easy to adjust the super parameter to control sparseness. Experimental results demonstrate that sparse controllable principal component analysis (SCPCA) is efficient.

Keywords Principal component analysis,Sparse penalty,Sparse controllable principal component analysis

1 引言

主成分分析(PCA)是一种无监督学习方法,通常用于数据降维处理[1]。PCA有着广泛的应用,比如手写邮政编码分类[2]、人脸识别[3]以及基因表达数据分析[4]。假设X是一个$n*m$的矩阵,表示一批数据有n条记录,且每条记录有m个属性,r是矩阵X的秩。PCA算法能顺序地找出r个荷载向量v,在荷载向量两两正交的条件下,使Xv的方差最大化[5]。利用PCA降低维度就是用线性变换提取最能保持数据特性的主要成分。在矩阵的属性具有实际的物理意义的情况下,对主成分做出解释是有必要的,但是这些分解得到的主成分通常是多个荷载向量的线性组合,而且主成分的荷载向量通常是非零值,在没有人为主观判断的情况下很难对主成分做出解释[12]。

为了克服PCA的这个缺点,到目前为止,相关文献包含了两大类解决方法:1)Jolliffe等提出的一系列的旋转技术,如Jolliffe和Uddin等提出的SCoT[6];2)在进行PCA时加上一个特定的约束项,即稀疏主成分分析,如Jolloiffe等提出的SCoTLASS[7],Zou等提出的SPCA[8]以及Shen等提出的用SVD实现稀疏主成分分析[5]。本文提出的稀疏可控主成分分析(SCPCA)算法属于上述中的第二类。由于引入了新的约束项及采用了稀疏可控投影算法,该算法具有易于调节稀疏参数和高效的特点。

本文第2节介绍稀疏化主成分分析的通用架构,并且提出稀疏可控约束项;然后将稀疏可控约束项运用到PCA,得到稀疏可控主成分分析算法(SCPCA),稀疏主成分分析算法的核心是一个稀疏可控投影;第3节给出一种将向量投影到给定稀疏水平的算法;第4节为实验部分;最后结合全文。

2 稀疏主成分分析

2.1 稀疏主成分分析的架构

X表示一个标准化的$n*m$的矩阵($\sum_{i=1}^{n}\boldsymbol{X}_{ij}=0,\sum_{i=1}^{n}X_{ij}^2=1$,

到稿日期:2015-12-26 返修日期:2016-04-14 本文受国家自然科学基金(61272274,60970063,31270101)资助。
谭亚芳(1990-),男,硕士,主要研究方向为机器学习、数据挖掘,E-mail:627484804@qq.com;刘 娟(1970-),女,教授,博士生导师,主要研究方向为生物信息处理、数据挖掘、机器学习;王才华(1987-),男,博士,主要研究方向为生物信息处理、机器学习;蒋万伟(1990-),男,硕士,主要研究方向为自然语言处理、数据挖掘。

$j=1,2,\cdots,m)$，X 的秩 $r\leqslant\min(n,m)$。那么 X 的第一个主成分求解过程可以表示为式(1)：

$$\max imize_v: v^{\mathrm{T}}X^{\mathrm{T}}Xv; \text{s. t. } \|v\|_2^2\leqslant 1, \|v\|_1\leqslant c \quad (1)$$

或者式(2)：

$$\operatorname{argmin}_{u,v}\|X-uv^{\mathrm{T}}\|_F^2; \text{s. t. } \|v\|_2^2\leqslant 1, \|v\|_1\leqslant c \quad (2)$$

其中，u 是维度大小为 n 的辅助列向量，v 是维度大小为 m 的列向量，c 是常量。式(1)是最常见的主成分分析形式，而式(2)与式(1)等价。

显然，式(2)中的可行解也是式(1)的可行解[9]。设 v^* 是式(2)的最优解，则有 $u^*=Xv^*$；$\|X-u^*v^{*\mathrm{T}}\|_2^2=\mathrm{Tr}(X^{\mathrm{T}}X)-v^{*\mathrm{T}}X^{\mathrm{T}}Xv^*$，所以式(1)和式(2)是等价的。注意到：式(2)的结论中有 $u^*=Xv^*$，即 u^* 为 v^* 对应的主成分。式(1)的求解问题不是凸优化问题，不易求解，而式(2)的求解过程是双凸优化问题，即固定其中一个变量对另一个变量的求解过程是一个凸优化问题。

利用式(2)可以求出第一个主成分。对于 X 的残差矩阵，重复迭代式(2)可求解出矩阵 X 的前 k(k 由使用者给定)个主成分。由此可以得到一个稀疏主成分分析的通用架构，其过程如算法 1 所示。

算法 1 稀疏主成分分析通用架构

1. 输入 X，需要提取主成分的个数 k
2. $X_1=X$
3. For i=1,2,…,k;
4. $u_i,v_i=\underset{u,v}{\operatorname{argmin}}\|X_i-uv^{\mathrm{T}}\|_F^2$ s. t. $u_k\perp u, 0<k<i$; $\|v\|_2^2=1$, P(v)$\leqslant$c
5. $X_{i+1}=X_i-u_iv_i^{\mathrm{T}}$
6. 返回$(u_1,\cdots,u_k),(v_1,\cdots,v_k)$

算法第 4 步是稀疏约束下矩阵 X 的 1 秩近似，第 5 步用于计算当前残差矩阵，重复迭代第 4－5 步 k 次，可输出主成分矩阵$(u_1,\cdots,u_k)$和荷载矩阵$(v_1,\cdots,v_k)$。普通的 PCA 算法不要求分解得到的主成分两两正交，但若得到的主成分是两两正交的，这对解释主成分是有益的[14]。所以第 4 步中加入了正交约束，$u_k\perp u, 0<k<i$。

2.2 PCA 目前已经使用的约束项

算法 1 第 4 步中的 $P(v)$称作向量 v 的约束项，参数 c 用来控制向量的稀疏度，传统通用的约束项如表 1 所列。

表 1 约束项函数

约束项	表达式
Lasso	$\|x\|_1$
Adaptive lasso	$\sum_{i=1}w_ix_i$
Fused lasso	$\sum_{i=1}x_i+\lambda\sum_{i=2}\|x_i-x_{i-1}\|$
SCAD	$\lambda\{I(x_i\leqslant\lambda)+\frac{(\alpha\|x_i\|-\lambda)}{(\alpha-1)\lambda}I(\|x_i\|>\lambda)\}$

Lasso 是一种常用的约束项，其优点是可以同时进行参数的估计和变量的选择[10]。adaptive lasso 是一种加权 lasso，对取值为 0 的数据系数有更高的权重[11]，SCAD 则可以自动一致地选取重要的变量，而且具有很高的效率。除了以上列出的常用约束项，还有多种约束项可以选择，比如 Xin Qi 等人提出的一种 L_1 正则和 L_2 正则混合使用的约束项，其可以较为容易地获得正交的荷载向量[13]。

2.3 稀疏可控约束项

表 1 列出的稀疏项虽然可以使主成分分析算法获得很好的稀疏性，但是稀疏参数(c)受向量的模(长度)、维度的影响，而且必须在一个非常大的区间内选取，所以稀疏参数很难确定。本文采用了一种新颖的基于 L_1 正则和 L_2 正则的约束项，如式(3)所示，该约束项可以克服上述缺点。

$$P_n(x)=\frac{\sqrt{n}-\frac{\|x\|_1}{\|x\|_2}}{\sqrt{n}-1} \quad (3)$$

其中，n 是 x 的维度大小，$P_n(x)\in(0,1)$表示稀疏水平，当 $P_n(x)\to 0$ 时，稀疏水平越低；当其等于 0 时，x 中的元素基本是非 0 的；$P_n(x)\to 1$ 时，稀疏水平越高，x 中为 0 的元素越多。

此约束项与其他约束项相比具有以下几个明显优点。

(1)对向量模不敏感：$\frac{\|\lambda x\|_1}{\|\lambda x\|_2}=\frac{\|x\|_1}{\|x\|_2}$，因此 $P_n(x)$与 x 的规模无关。

(2)对向量的维度不敏感：$P_n(x)$在分子分母上用到$\sqrt{n}$，所以此约束项对 x 的维度不敏感。

(3)约束项取值范围在 0～1 之间，$P_n(x)\in[0,1]$使得稀疏参数选择较为容易。

式(3)具有以上的特性，它能很好地控制向量的稀疏水平，称之为稀疏可控约束项。在处理相关问题时，如果有相关的背景知识，可以极大地削减参数的取值范围，而对于其他稀疏约束项，这是很难做到的，因为 $P_n(x)$与向量的规模和维度有关。

2.4 稀疏可控主成分分析

稀疏可控主成分分析(SCPCA)采用算法 1 的架构，区别在于将第 4 步变为式(4)所示的优化问题：

$$\begin{aligned}&\min_{u,v}\|X-uv^{\mathrm{T}}\|_F^2\\&\text{s. t. } \|v\|_2^2=1, P(v)\geqslant c; u_i\perp u_j, i\neq j, i,j=1,2,\cdots,k\end{aligned} \quad (4)$$

其中，$P_n(v)=\frac{\sqrt{n}-\frac{\|v\|_1}{\|v\|_2}}{\sqrt{n}-1}$，$n$ 是向量 v 的维度，$c\in[0,1]$，F 表示目标矩阵的 Frobenious 范数，目标函数 $\min_{u,v}\|X_i-uv^{\mathrm{T}}\|_F^2$对于 u,v 是一个双凸优化问题，即固定其中一个变量，目标函数对于另一个变量的求解是一个凸优化问题。

在式(4)中，首先假设 u 固定，设 $z=\frac{u^{\mathrm{T}}X}{\|u\|_2^2}$，则有：

$$\begin{aligned}\|X-uv^{\mathrm{T}}\|_F^2&=\mathrm{Tr}((X-uv^{\mathrm{T}})^{\mathrm{T}}(X-uv^{\mathrm{T}}))\\&=\|u\|_2^2*\|v\|_2^2-2u^{\mathrm{T}}Xv+\mathrm{Tr}(X^{\mathrm{T}}X)\\&=\|u\|_2^2*\|v-z\|_2^2+C\end{aligned}$$

其中 $C=\mathrm{Tr}(X^{\mathrm{T}}X)-\|z\|_2^2*\|u\|_2^2$，显然式(4)在求解 v 时可以改写为式(5)：

$$\min_v\|v-z\|_2^2; \text{s. t. } \|v\|_2^2=1; P(v)\geqslant c \quad (5)$$

式(5)的求解算法将在本文的第 3 节给出。在求解得到 v_k(第 k 次求解 v 时)之后，则 v_k 已知，令 $u=du'$，$\|u'\|_2^2=1$，则式(4)求解 u 变为式(6)：

$$\min_{d,u'}\|X_k-du'v_k^{\mathrm{T}}\|_F^2; \text{s. t. } \|u'\|_2^2=1, u_k^2\perp u_1,\cdots,u_{k-1} \quad (6)$$

记矩阵 U_{k-1} 表示已经求解得到的$(u_1,\cdots,u_{k-1})$对应的单位向量$(u_1',\cdots,u_{k-1}')$，用 $U_{k-1}^{\perp}$(未知)表示一个与 U_{k-1} 正交的矩阵，且 $U_{k-1}^{\perp}$ 中的列向量为单位向量，则有 $U=(U_{k-1},U_{k-1}^{\perp})$，$UU^{\mathrm{T}}=U^{\mathrm{T}}U=I$。由于 u_k' 与 U_{k-1} 中向量正交，u' 可以

用 $U_{k-1}^{\perp}$ 中的向量线性表示。假设 $u'=U_{k-1}^{\perp}\theta$，则式(6)等价为式(7)：

$$\min_{d,\theta}: \| X_k - dU_{k-1}^{\perp}\theta v_k^{\mathrm{T}} \|_F^2 \tag{7}$$

可以得到 θ, d 的最优解如式(8)所示：

$$\theta=\frac{U_{k-1}^{\perp\mathrm{T}}Xv_k}{\| U_{k-1}^{\perp\mathrm{T}}Xv_k \|_2}, d=\sqrt{v_k^{\mathrm{T}}X^{\mathrm{T}}(I-U_{k-1}U_{k-1}^{\mathrm{T}})Xv_k} \tag{8}$$

因此可以得到 u_k 的解为式(9)：

$$u_k=dU_{k-1}^{\perp}\theta=\frac{U_{k-1}^{\perp}U_{k-1}^{\perp\mathrm{T}}Xv_k}{\| U_{k-1}^{\perp\mathrm{T}}Xv_k \|_2}=(I-U_{k-1}U_{k-1}^{\mathrm{T}})Xv_k \tag{9}$$

随着 u, v 被确定，式(4)的求解可以描述为算法 2。

算法 2　稀疏主成分分析 1 秩近似算法

1. 输入矩阵 X，稀疏控制常量 c，正交矩阵 U_{k-1}
2. 初始化向量 u
3. 迭代步骤 4－步骤 6 直到收敛
4. $z=u^{\mathrm{T}}X/\| u \|_2^2$
5. $v=\mathrm{argmin}_v \| v-z \|_2^2, \| v \|_2^2=1, P_n(v)\geqslant c$
6. $u=(I-U_{k-1}U_{k-1}^{\mathrm{T}})Xv$
7. 返回 u，v

算法 2 总结了稀疏可控主成分分析(SCPCA)中矩阵的 1 秩近似过程，算法的输入为一个矩阵 X 和一个稀疏水平控制常量 c，返回向量 u 和稀疏向量 v，轮流确定 u, v 的值，确定 u 之后，用稀疏可控投影算法计算 v，相关投影算法将在第 3 节中讨论。

3　稀疏可控投影算法

3.1　基本的稀疏可控投影算法

稀疏可控投影算法把一个向量投影到指定的稀疏水平，算法可以描述为式(10)的优化问题：

$$\min_x \frac{1}{2} \| x-z \|_2^2; \text{s. t. } P_n(x)\geqslant c \tag{10}$$

其中，x, z 是向量，n 是 x 的维度大小，$c\in[0,1]$，$P_n(x)=\frac{\sqrt{n}-\frac{\| x \|_1}{\| x \|_2}}{\sqrt{n}-1}$，设 $m=(1-c)\sqrt{n}+c$，则 $\frac{\| x \|_1}{\| x \|_2}\leqslant m$，引入变量 θ，使得 $\theta=\| x \|_2^2$，则式(10)可转化为式(11)：

$$\min_x \frac{1}{2} \| x-z \|_2^2; \text{s. t. } \| x \|_2^2=\theta, \| x \|_1\leqslant\sqrt{\theta}m \tag{11}$$

引理 1　如果 x^* 是式(11)的最优解，那么对于所有的 i，有 $z_i x_i^*\geqslant 0$。

证明：已知 x^* 是式(11)的一个最优解，若存在一个 i 使得 $z_i x_i^*\leqslant 0$，构造一个向量 y 使得 $y_i=-x_i^*$，当 $j\neq i$ 时 $y_i=x_i^*$，因此有 $\| y \|_1=\| x^* \|_1$，$\| y \|_2=\| x^* \|_2$，$\| x^*-z \|_2^2-\| y-z \|_2^2=2(z_i y_i-z_i x_i^*)=-4z_i x_i^*$，由此可得 y 也是式(11)的一个解，这与题设矛盾。

根据引理 1 可以削减解的空间，让 $|z|$ 表示取 z 向量每个元素绝对值之后形成的向量，则式(11)可以改写为式(12)：

$$\min_x \frac{1}{2} \| x-|z| \|_2^2;$$
$$\text{s. t. } \| x \|_2^2=\theta, \sum_{i=1}^n x_i\leqslant\sqrt{\theta}m, x_i\geqslant 0, i=1,2,\cdots,n \tag{12}$$

引理 1 表明如果 x^* 是式(12)的解，则 $\mathrm{sign}(z)x^*$ 是式(11)的解，sign()是符号函数。求解式(12)的拉格朗日函数如式(13)所示：

$$L(x)=\frac{1+\lambda}{2}x^{\mathrm{T}}x-|z|^{\mathrm{T}}x+\eta\sum_{i=1}^n x_i-\sum_{i=1}^n\zeta_i x_i+C' \tag{13}$$

其中，λ 是实数，$\eta\geqslant 0, \zeta_i\geqslant 0$ 是拉格朗日系数，C' 对于向量 x 是常量，若 $1+\lambda<0$，则式(12)无最小值，这与求解目标相违背；若 $1+\lambda=0$，则式(13)是线性的，在 $x=0$ 时取得最小值，表明向量中的元素全为 0，此时由于 $\frac{\| x \|_1}{\| x \|_2}$ 条件不满足，因此 $1+\lambda>0$。对式(13)中 x 求导数得到式(14)：

$$\frac{\partial L(x)}{\partial x_i}=(1+\lambda)x_i-|z_i|+\eta-\zeta_i=0 \tag{14}$$

而式(14)的解如式(15)所示：

$$x^*=\frac{\max(|z|-\eta,0)}{1+\lambda} \tag{15}$$

由于式(15)中 η, λ 是未知的拉格朗日系数，但是 x 满足条件 $\| x \|_2^2=\theta$，$\| x \|_1\leqslant\sqrt{m}\theta$，因此消去变量 θ 则可以得到一个如式(16)所示的关于 η 的不等式：

$$f(\eta)=\frac{\| x^* \|_1}{\| x^* \|_2}=\frac{\| \max(|z|-\eta,0) \|_1}{\| \max(|z|-\eta,0) \|_2} \tag{16}$$

如果 $\eta=0$，$\frac{\| x^* \|_1}{\| x^* \|_2}\leqslant m$，则 $|z|$ 是式(12)的解，否则需要找到一个满足 $f(\eta)=m$ 的 η，假设 η^* 是 $f(\eta)=m$ 的一个最优解，η^* 的求解算法将在 3.3 节给出，引入变量 $\hat{x}$ 表示 $\max(|z|-\eta^*,0)$，由式(15)可得 $x^*=\left(\frac{|z|^{\mathrm{T}}\hat{x}}{\hat{x}^{\mathrm{T}}\hat{x}}\right)\hat{x}$，其中 $\lambda=\frac{\hat{x}^{\mathrm{T}}\hat{x}}{|z|^{\mathrm{T}}\hat{x}}-1$，由引理 1 可得式(11)的最优解 $\mathrm{sign}(z)x^*=\mathrm{sign}(z)\left(\frac{|z|^{\mathrm{T}}\hat{x}}{\hat{x}^{\mathrm{T}}\hat{x}}\right)\hat{x}$。

算法 3　稀疏可控投影算法

1. 输入维度大小为 n 的向量 z、稀疏水平 c
2. 初始化 $m=(1-c)\sqrt{n}+c$
3. 如果 $\frac{\| z \|_1}{\| z \|_2}\leqslant m$，返回 z
4. 找到一个 η^* 使得 $f(\eta)=m$，$\hat{x}=\max(|z|-\eta^*,0)$ 成立
5. 返回 $\mathrm{sign}(z)\left(\frac{|z|^{\mathrm{T}}\hat{x}}{\hat{x}^{\mathrm{T}}\hat{x}}\right)\hat{x}$

算法 3 接受一个维度大小为 n 的向量和一个稀疏水平为 c 的常量作为输入，输出一个维度大小为 n 的稀疏的向量，此算法描述了稀疏可控投影算法的过程，算法中 η^* 是未知的，3.2 节、3.3 节将介绍 $f(\eta)$ 函数的特性，然后设计一个有效的算法来得到 η 的最优解。

3.2　$f(\eta)$ 函数是一个不增函数

本节将证明 $f(\eta)$ 函数在其定义域内是一个不增的函数，为方便起见，用 $z_{(i)}$ 表示向量 $|z|$（假设 $|z|$ 是一个降序排列的向量）中第 i 个元素，即 $z_{(1)}\geqslant z_{(2)}\geqslant\cdots\geqslant z_{(n)}\geqslant 0, z_{(i)}\in\mathcal{R}$。引入一个如式(17)所示的辅助函数：

$$f_k(\eta)=\frac{\sum_{i=1}^k(z_{(i)}-\eta)}{\sum_{i=1}^k(z_{(i)}-\eta)^2}, \eta\in(z_{(k+1)}, z_{(k)}] \tag{17}$$

$f_k(\eta)$ 函数的导数（函数定义域两端的导数在这里不讨论，参见式(20)、式(21)）为：

$$f_k'(\eta)=-\frac{k^2\left(\frac{1}{k}\sum_{i=1}^{k}(z_{(i)}-\eta)^2-\left(\frac{1}{k}\sum_{i=1}^{k}(z_{(i)}-\eta)\right)^2\right)}{\sqrt{\left(\sum_{i=1}^{k}(z_{(i)}-\eta)^2\right)^{\frac{3}{2}}}} \tag{18}$$

在统计学中随机变量 x 的方差可以用 $D(x)=E(x^2)-E^2(x)$ 计算，且 $D(x)\geqslant 0$，所以式(18)中 $f_k'(\eta)\leqslant 0$，对于每个固定的 k 值，$f_k(\eta)$ 函数在其定义域内是一个不增函数。

辅助函数 $f_k(\eta)$ 和 $f(\eta)$ 函数的关系可以描述为式(19)：

$$f(\eta)=\frac{\|\max(|z|-\eta,0)\|_1}{\|\max(|z|-\eta,0)\|_2}=f_k(\eta),\eta\in(z_{(k+1)},z_{(k)}] \tag{19}$$

由于 $f(\eta)$ 是一个分段函数，接下来探讨 $f(\eta)$ 函数在点 $z_{(k+1)}$ 处的相关性质。在临界点 $z_{(k+1)}$ 处，$f(\eta)$ 函数的左右极限分别如式(20)、式(21)所示：

$$\lim_{\eta\to z_{k+1}^-} f_k(\eta)=\frac{\sum_{i=1}^{k}(z_{(i)}-z_{(k+1)})}{\sqrt{\sum_{i=1}^{k}(z_{(i)}-z_{(k+1)})^2}} \tag{20}$$

$$\lim_{\eta\to z_{k+1}^+} f_{k+1}(\eta)=\frac{\sum_{i=1}^{k+1}(z_{(i)}-z_{(k+1)})}{\sqrt{\sum_{i=1}^{k}(z_{(i)}-z_{(k+1)})^2}} \tag{21}$$

由此可得 $f(\eta)$ 函数在临界点 $z_{(k+1)}$ 处左右极限相等，而且对于每个固定的 k 值，$f_k(\eta)$ 函数在其定义域内是一个不增函数，进而可得 $f(\eta)$ 函数在取不同的 k 值时连续，所以 $f(\eta)$ 函数在其定义域内是一个单调连续不增函数。

3.3 一个快速得到 η 的算法

因 $f(\eta)$ 函数是连续单调不增的函数，要找出满足 $f_k(\eta)=m$ 条件的解有多种方法，常规的方法有：

1)线性搜索、二分查找。搜索到一个合适的 η 满足条件 $|f(\eta)-m|<\varepsilon$ 为止。

2)随机搜索算法。找到一个合适的 k 使得 $f_{k+1}(z_{(k+1)})\leqslant m\leqslant f_k(z_{(k+1)})$，然后解方程 $f_k(\eta)=m$。

本文采用随机算法求解 η，随机算法如算法 4 所示。

算法 4 随机搜索算法

1. 输入 n 维度大小的向量 z 和稀疏水平 c
2. 初始化 U=[n]，s1=0；s2=0，k=0
3. 当 U 不为空时
4. 在 U 中随机选取一个 k
5. 按以下规则划分 U：G={j∈U|$z_{(j)}\geqslant z_{(k)}$} L={j∈U|$z_{(j)}<z_{(k)}$}
6. 计算 $ds1=\sum_{j\in G}z_{(j)}$，$ds2=\sum_{j\in G}z_{(j)}^2$，dk=|G|，ts1=s1+ds1，ts2=s2+ds2，tk=k+dk
7. IF $\frac{ts1-tk*z_{(k)}}{ts2-2ts1*z_{(k)}+tk*z_{(k)}^2}<m$，s1=ts1，s2=ts2，k=tk，U=L
8. ELSE U=G−{k}
9. $r1=s1/k,r2=s2/k,\eta=r1-\sqrt{\frac{m^2}{k-m^2}}(r2-r1^2)$
10. 返回 η

4 实验结果

4.1 稀疏可控主成分分析对稀疏的控制

引言中提到，经 SCSCP 算法分解后的矩阵的稀疏度可以达到指定的水平。随机生成一个维度大小为 100 的方阵，其中元素的取值范围为 0～100，提取主成分的个数为 30，本文实验平台为：CPU：Inter（R）Core（TM）i3-2102 3.3GHz；RAM：6GB；OS：Windows 7 64bit 。

表 2 稀疏可控效果

稀疏参数 c 取值	实际稀疏度
0.10	0.007
0.20	0.100
0.30	0.180
0.40	0.330
0.50	0.480
0.60	0.630
0.70	0.760
0.80	0.850
0.90	0.930
0.99	0.980

表 2 中参数 c 是算法 3 中的稀疏水平参数，实际稀疏度是得到的主成分荷载向量中值为 0 的元素的平均数量(3 次结果的平均值)所占总量的比例。表 2 表明：在稀疏度特别低时，结果与指定的稀疏度有所差别，c 取值为 0.2 时差别已不明显；到 0.4 以上时，实际的稀疏度和指定的稀疏度基本相符，实验表明稀疏可控算法可以将向量投影到指定的水平，而且稀疏度的取值范围在 0～1。由于 SCSCP 分解得到的荷载向量接近指定的稀疏水平，因此本算法比其他稀疏主成份分析算法更易于控制稀疏性。

4.2 稀疏可控算法的效率

表 3 中第一列是随机生成方阵维度的大小，第一行是稀疏主成分分析算法采用的约束项，第 2、3 列用的约束项都是 lasso，不同的是 lars 使用最小角回归方法，而 cd 使用梯度下降方法，scp 即稀疏可控主成分分析，实验中提取主成分个数为 30，误差为 1E−8，实验结果都为平均值，单位为秒。

表 3 稀疏可控算法的效率

维度	lars	cd	scp
100	51.46	1.13	0.06
200	102.30	2.51	0.34
300	152.34	3.69	0.95
400	226.70	12.96	2.36
500	262.49	13.43	4.82
600	241.51	18.45	8.49
700	369.44	26.03	12.50
800	427.98	27.55	17.23
900	504.33	34.11	22.27
1000	534.54	40.44	27.91

从表 3 中可以看出，随着矩阵维度大小的增加，3 种算法所用时间增加迅速；用最小角回归方法求解带 lasso 约束项的主成分分析算法的速度明显慢于其他两者，而 SCSCP 算法由于采用了稀疏可控投影算法，速度也要明显快于带 lasso 约束项的稀疏主成分分析算法。3 种算法提取一次主成分所用的时间复杂度都为 $O(n^3)$ 级，稀疏惩罚项(scp)采用的随机搜索算法与二分查找算法相比，虽然都是 $O(\log n)$ 级，但是二分查找算法需要在一个巨大的实数范围内搜索方程的解，因此引入稀疏惩罚项的稀疏可控主成分分析是高效的。

结束语 本文在稀疏主成分分析时引入了稀疏可控约束项，这种约束项具有规模和维度不敏感的特性，而且稀疏水平的值在 0～1 之间。这些特性使得对稀疏性的控制变得容易而且稀疏可控投影算法极大地提高了稀疏可控主成分分析算法的效率。

参 考 文 献

[1] JOLLIFFE I T. Principal Component Analysis(second ed.)[M]. New York, Springer, 2002.

(下转第 282 页)

ment,2014,51(4):691-706. (in Chinese)
丁兆云,贾焰,周斌. 微博数据挖掘研究综述[J]. 计算机研究与发展,2014,51(4):691-706.

[7] LI D,NIU J,QIU M,et al. Sentiment analysis on Weibo data [C]// 2014 IEEE Computing,Communications and IT Applications Conference (ComComAp). IEEE,2014:249-254.

[8] LIAN Jie,ZHOU Xin,LIU Yun. SINA microblog data retrieval [J]. J T sing hua Univ(Sci & Tech),2011,51(10):1300-1305. (in Chinese)
廉捷,周欣,刘云. 新浪微博数据挖掘方案[J]. 清华大学学报(自然科学版),2011,51(10):1300-1305.

[9] HUANG Yan-wei,LIU Jia-yong. Study on Sinamicroblog Data Acquisition Technology[J]. Information Security and Communication Security,2013(6):71-73. (in Chinese)
黄延炜,刘嘉勇. 新浪微博数据获取技术研究[J]. 信息安全与通信保密,2013(6):71-73.

[10] YAO Ke. Open API: Sina micro Bo way? [J]. Internet World, 2010(8):71-72. (in Chinese)
姚科. 开放 API:新浪微博必经之路? [J]. 互联网天地,2010(8):71-72.

[11] LI X,XIE Y,LI C,et al. Analyzing the public events' influence via open microblogging APIs[C]// 2012 International Conference on Machine Learning and Cybernetics (ICMLC). IEEE, 2012:84-90.

[12] SUN Xiao,YE Jia-qi,TANG Chen-yi,et al. Method of Sina microblogging big data grabbing based on multi-strategy and its application[J]. Journal of Hefei University of Technology, 2014,37(10):1210-1215. (in Chinese)
孙晓,叶嘉麒,唐陈意,等. 基于多策略的新浪微博大数据抓取及应用[J]. 合肥工业大学学报(自然科学版),2014,37(10):1210-1215.

[13] YAO Feng. Improvement of Base64 Encoding/Decoding Algorithm in Java[J]. Computer Applications and Software,2008,25 (12):164-165. (in Chinese)
姚峰. Java 平台中 Base64 编码/解码算法的改进[J]. 计算机应用与软件,2008,25(12):164-165.

[14] SUN Qing-yun,WANG Jun-feng,ZHAO Zong-qu,et al. A Microblog Data Collection Method Based on Simulated Login Technology[J]. Computer Technology and Development, 2014, 24 (3):6-10. (in Chinese)
孙青云,王俊峰,赵宗渠,等. 一种基于模拟登录的微博数据采集方案[J]. 计算机技与发展,2014,24(3):6-10.

[15] DANGRE A,WANKHEDE V,AKRE P,et al. Design and Implementation of Web Crawler[J]. International Journal of Computer Science & Information Technolo,2014,5(1):921-922.

[16] SHEN D,WANG H,CAO J,et al. The Design and Implement of High Efficient Incremental Microblogging Crawler[C]// 2012 Fourth International Conference on Multimedia Information Networking and Security (MINES). IEEE,2012:537-540.

[17] VASILE A I,PAVALOIU B,CRISTEA P D. Building a specialized high performance web crawler[C]// 2013 20th International Conference on System,Signals and Image Processing (IWSSIP). IEEE,2013:183-186.

[18] WANG Ye. The design and implementation of the theme crawler based on the breadth first[D]. Shanghai: Fudan University, 2011. (in Chinese)
王桦. 基于广度优先的主题爬虫的设计与实现[D]. 上海:复旦大学,2011.

[19] LIAN Jie. Research on social network data mining based on user characteristics [D]. Beijing: Beijing Jiaotong University, 2013. (in Chinese)
廉捷. 基于用户特征的社交网络数据挖掘研究[D]. 北京:北京交通大学,2013.

[20] LIU J,CAO Z,CUI K,et al. Identifying Important Users in Sina Microblog[C]// 2012 Fourth International Conference on Multimedia Information Networking and Security (MINES). IEEE, 2012:839-842.

(上接第 246 页)

[2] HASTIE T,TIBSHIRANI R,FRIEDMAN J. The Elements of Statistical Learning [M]. Data mining,Interface and Prediction New York,Springer,2001.

[3] HANCOCK P J B,BURTON A M,BRUCE V. Face processing: human perception and principal components analysis [J]. Memory and Cognition,1996,24(1):26-40.

[4] MISRA J,SCHMITT W,et al. Interactive Exploration of Microarray Gene Expression Patterns in a Reduced Dimensional Space [J]. Genome Research,2012,12(7):1112-1120.

[5] SHEN Hai-peng,HUANG Jian-hua. Sparse principal component analysis via regularized low rank matrix approximation [J]. Journal of Multivariate Analysis,2008,99(6):1015-1034.

[6] JOLLIFFE I T,UDDIN M. The Simplified Component Technique: An Alternative to Rotated Principal Components [J]. Journal of Computational and Graphical Statistics,2000,9(9): 689-710.

[7] JOLLIFFE I T,TRENDAFILOV N T,et al. A Modified Principal Component Technique Based on the LASSO [J]. Journal of Computational and Graphical Statistics,2003,12(3):531-547.

[8] ZOU H,HASTIE T,et al. Sparse principal component analysis [J]. Journal of Computational and Graphical Statistics, 2006, 15:265-286.

[9] WITTEN D M,et al. A penalized matrix decomposition,with applications to sparse principal components and canonical correlation [J]. Biostatistics,2009,10(3):515-534.

[10] TIBSHIRANI R. Regression shrinkage and selection via the lasso [J]. Journal of the Royal Statistical Society, 1996, 58(1): 267-288.

[11] ZOU H. The adaptive lasso and its oracle properties [J]. Journal of the American Statistical Association, 2006, 101(476): 1418-1429.

[12] ALLEN G I,GROSENICK L,et al. the A Generalized Least-Square Matrix Decomposition [J]. Journal of the American Statistical Association,2014,109(505):145-159.

[13] QI Xin,LUO Rui-yan,ZHAO Hong-yu. Sparse principal component analysis by choice of norm[J]. Journal of Multivariate Analysis,2013,114(2):127-160.