# 基于迹比率的多数据集判别分析维数压缩

赵小彤[1]，李　智[3]，宋恩彬[1,2]
(1. 四川大学数学学院，成都 610064；2. 四川大学空天科学与工程学院，成都 610064；3. 电子信息控制重点实验室，成都 610063)

**摘　要**：主成分分析法常被用于维数压缩和特征提取，其在处理单一高维数据集时有很大优势. 在很多实际场景中需要联合处理多个数据集，此时传统的主成分分析方法面临很大挑战. 本文提出了迹比率主成分分析法，该方法可以提取目标数据相对其他数据特有的低维表示，进而通过迭代算法高效求解. 数值算例证实了该方法的优越性.

**关键词**：主成分分析；判别分析；维数压缩；多背景数据集
**中图分类号**：O29　　**文献标识码**：A　　DOI：10.19907/j.0490-6756.2021.011002

# Trace ratio-based dimensionality reduction for discriminative analysis of multiple datasets

*ZHAO Xiao-Tong*[1]，*LI Zhi*[3]，*SONG En-Bin*[1,2]
(1. School of Mathematics, Sichuan University, Chengdu 610064, China;
2. School of Aeronautics & Astronautics, Sichuan University, Chengdu 610064, China;
3. Science and Technology on Electronic Information Control Laboratory, Chengdu 610063, China)

**Abstract**: Principal component analysis is widely applied in dimensionality reduction and feature extraction, especially in tackling single high-dimensional dataset. However, traditional principal component analysis faces challenge when it comes to analyzing multiple datasets jointly. This paper introduces a novel approach named trace ratio principal component analysis, which can discover low-dimensional structure unique to the target data relative to others. Furthermore, trace ratio principal component analysis and its variants can be solved by efficient iterative algorithm. Numerical experiments show the efficiency of the method.

**Keywords**: Principal component analysis; Discriminative analysis; Dimensionality reduction; Multiple background datasets

## 1　引　言

主成分分析算法(PCA)是数据挖掘和特征提取的有效工具，在生物信息学[1-2]、基因组学、定量金融学和信号处理等领域有广泛应用. PCA 的目标是得到高维数据的低维表示，同时尽可能保留高维数据的信息[3]. 然而，很多实际场景通常涉及到多个数据集，其中人们感兴趣的是提取出某个数据集相对其他数据集特有的信息[1-17]. 例如，考虑两个基因表达观测数据集，它们来自于对多个种族人群的观测. 第一个数据集包含癌症患者的基因表达，是我们感兴趣的数据集，称之为目标数据. 第二个数据集则是由健康人群的基因表达组成，称为背景数据. 如果将 PCA 单独应用于目标数据，或

**收稿日期**：2019-05-29
**基金项目**：四川省科技计划项目(2019YJ0115)；四川大学基金(2020SCUNG205)；国家自然科学基金(U2066203，61473197)
**作者简介**：赵小彤(1995－)，男，河北秦皇岛人，硕士生，主要研究方向为应用统计与优化理论. E-mail：zhaoxiaotong6@163.com
**通讯作者**：宋恩彬. E-mail：e.b.song@163.com

目标数据与背景数据的结合,可能都仅能获得两个数据集公共的背景信息(人口模型、性别等),而不能描述癌症患者的判别信息. 为处理上述问题,Abid 等[4]首次提出了对比主成分分析算法(cPCA),并将其用于提取两个数据集之间的判别信息. 具体而言,cPCA 寻找使得目标数据方差极大且背景数据方差极小的方向. cPCA 在选取超参数时需要进行多次特征分解和谱聚类算法[6],计算复杂度高. 基于文献[4],文献[5]给出了判别主成分分析算法(dPCA),dPCA 不需要额外的超参数,计算复杂度会低一些. 但 dPCA 不能处理背景方差矩阵奇异的情形,且进行后续分类算法时分类效果不佳,对高维数据的低维特征提取合理性有待改进. 近期,很多学者将这两种算法应用于各种实际场景[7-8],但这两种算法在实际应用中都具有一定的局限性.

本文提出了一种处理多数据集的新方法,称为迹比率主成分分析算法(trPCA). trPCA 通过求解一个迹比率问题提取出目标数据相对背景数据特有的低维表示. trPCA 能够处理背景方差阵奇异的情形,并且计算成本低于 cPCA. 进一步,针对多个背景数据集情形,本文将 trPCA 推广给出了多背景数据集迹比率主成分分析算法(MtrPCA),该方法可以提取出某个数据集相对其他所有背景数据集特有的低维表示.

## 2　预备知识

考虑两个数据集,即目标数据集 $\{x_i\}_{i=1}^n$ 和背景数据集 $\{y_j\}_{j=1}^m$,其中 $x_i, y_j \in \mathbf{R}^d$. 我们要寻找目标数据而不是背景数据具有的特征. 不失一般性,假设两个数据集都已被中心化,即已从每个 $x_i$(或 $y_j$)中减去了样本均值 $\frac{1}{n}\sum_{i=1}^n x_i$(或 $\frac{1}{m}\sum_{j=1}^m y_j$). 接下来我们先概述 PCA,cPCA 以及 dPCA 算法.

PCA 算法每次处理一个数据集. 它通过线性投影得到高维数据的低维表示. 具体而言,投影矩阵 $\hat{U} \in \mathbf{R}^{d\times k}$ 由以下模型的最优解给出:

$$\max_{U\in\mathbf{R}^{d\times k}} \operatorname{tr}(U^{\mathrm{T}} C_{xx} U),\ \text{s.t.}\quad U^{\mathrm{T}}U = I_k \tag{1}$$

其中 $C_{xx} := \frac{1}{n}\sum_{i=1}^n x_i x_i^{\mathrm{T}} \in \mathbf{R}^{d\times d}$ 是目标数据集 $\{x_i\}_{i=1}^n$ 的样本协方差矩阵,称为目标方差. 问题(1)的最优解 $\hat{U}$ 的 $k$ 个列向量即为 $C_{xx}$ 的前 $k$ 个最大特征值对应的标准正交特征向量,相应的投影 $\{z_i\} = \{\hat{U}^{\mathrm{T}} x_i\}_{i=1}^k$ 构成目标数据的前 $k$ 个主成分(PCs). 无论将 PCA 单独应用于 $\{x_i\}_{i=1}^n$ 或 $\{\{x_i\}_{i=1}^n, \{y_j\}_{j=1}^m\}$,我们都不能提取出目标数据相对于背景数据具有判别的特征.

文献[4]提出的 cPCA 方法旨在寻找子空间使得目标数据集方差尽量大且背景数据集方差尽量小. 具体而言,我们解如下问题:

$$\max_{U\in\mathbf{R}^{d\times k}} \operatorname{tr}[U^{\mathrm{T}}(C_{xx} - \alpha C_{yy})U],\quad \text{s.t. } U^{\mathrm{T}}U = I_k \tag{2}$$

其中 $C_{yy} := \frac{1}{m}\sum_{j=1}^m y_j y_j^{\mathrm{T}} \in \mathbf{R}^{d\times d}$ 是背景数据集 $\{y_j\}_{j=1}^m$ 的样本协方差矩阵,称为背景方差,超参数 $\alpha \in [0, \infty]$ 表示极大化目标方差与极小化背景方差之间的权衡系数. 对于给定的 $\alpha$,问题(2)求得 $C_{xx} - \alpha C_{yy}$ 的前 $k$ 个最大特征值对应的标准正交特征向量张成的子空间,然后将 $x_i$ 向这个子空间投影构成数据集的前 $k$ 个对比主成分(cPCs).

当 $\alpha = 0$ 时,cPCA 只选择极大化目标方差的方向,从而退化为只对 $x_i$ 做 PCA. 当 $\alpha = \infty$ 时,cPCA 将目标数据投影到背景数据方差阵的零空间. 因此,$\alpha$ 的选取至关重要. 在文献[4]中,cPCA 先根据多个 $\alpha$ 的预选值分别进行特征分解,然后再进行谱聚类,最终输出几个 $\alpha$ 值以及相应的子空间. 但是,对于高维数据,该算法会产生很高的计算复杂度,并且选择不适当的 $\alpha$ 会导致结果有很大偏差.

基于文献[4],文献[5]提出了 dPCA 方法,即求解下述比率迹(Ratio Trace)问题:

$$\max_{U\in\mathbf{R}^{d\times k}} \operatorname{tr}[(U^{\mathrm{T}} C_{yy} U)^{-1} U^{\mathrm{T}} C_{xx} U] \tag{3}$$

该问题等价于寻求 $C_{yy}^{-1} C_{xx}$ 的前 $k$ 个最大特征值对应的特征子空间[9]. 这种方法不需要额外的超参数,计算复杂度低于 cPCA 算法,但该算法要求 $C_{yy}$ 可逆,无法应用于 $C_{yy}$ 不可逆情形,并且问题(3)的最优解在进行一个非奇异变换之后仍是最优解,从而可能会对后续的分类、聚类问题产生影响. 因此,很多情形下 dPCA 算法得到的低维特征表示效果不佳.

## 3　算法描述

### 3.1　迹比率主成分分析

基于上述分析,我们提出一种新方法. 考虑如下的迹比率(Trace Ratio)问题,我们称之为迹比率主成分分析算法(trPCA):

$$\max_{U\in\mathbf{R}^{d\times k}}\frac{\mathrm{tr}(U^{\mathrm{T}}C_{xx}U)}{\mathrm{tr}(U^{\mathrm{T}}C_{yy}U)},\ \text{s.t. } U^{\mathrm{T}}U=I_k \tag{4}$$

该问题尽管没有解析解，但已有大量的迭代算法可以快速求解[10-12]. 记问题(4)的最优解为 $\hat{U}$. 我们称$\{\hat{U}^{\mathrm{T}}x_i\}_{i=1}^{k}$为目标数据的迹比率主成分(trPCs)，其中 $k$ 表示所提取的主成分维数，可以根据经验事先给定或分别对多个 $k$ 的预选值进行降维算法选择后续分类或聚类效果好的 $k$. 利用文献[10]中的二分法求解问题(4)，该算法具有全局最优收敛性，具体执行步骤如下.

**算法 3.1**　迹比率主成分分析

1. 输入：目标数据$\{x_i\}_{i=1}^{n}$、背景数据$\{y_j\}_{j=1}^{m}$和提取的主成分维数 $k$；

2. 将$\{x_i\}_{i=1}^{n}$和$\{y_j\}_{j=1}^{m}$中心化；

3. 计算经验协方差矩阵 $C_{xx}$和$C_{yy}$；

4. 用二分法[10]求解问题(4)；

5. 输出：子空间 $\hat{U}\in\mathbf{R}^{d\times k}$.

值得注意的是，虽然很多监督、半监督维数压缩算法最终都会转为求解迹比率问题，如线性判别分析(LDA)[13]、边界 Fisher 分析(MFA)[14]、局部线性嵌入(LLE)等，但这些算法所解决的问题都与本文不同.

通过下面的定理，我们可以得出 trPCA 和 cPCA 之间的关系，证明详见文献[8].

**定理 3.2**　对 $d$ 阶实对称矩阵 $C_{xx}\geq 0$ 和 $C_{yy}>0$，$U\in\mathbf{R}^{d\times k}$，

$$\hat{\alpha}=\frac{\mathrm{tr}(\hat{U}^{\mathrm{T}}C_{xx}\hat{U})}{\mathrm{tr}(\hat{U}^{\mathrm{T}}C_{yy}\hat{U})}=\max_{U^{\mathrm{T}}U=I_k}\frac{\mathrm{tr}(U^{\mathrm{T}}C_{xx}U)}{\mathrm{tr}(U^{\mathrm{T}}C_{yy}U)}$$

当且仅当

$$0=\mathrm{tr}[\hat{U}^{\mathrm{T}}(C_{xx}-\hat{\alpha}C_{yy})\hat{U}]=\max_{U^{\mathrm{T}}U=I_k}\mathrm{tr}[U^{\mathrm{T}}(C_{xx}-\hat{\alpha}C_{yy})U],$$

其中 $\hat{U}\in\mathbf{R}^{d\times k}$为极大化问题的最优解.

由定理 3.2 可知，当 $\alpha$ 等于某个值时，trPCA 和 cPCA 具有一定的等价性.

前文已经提到，dPCA 的解经过任何一个非奇异矩阵变换之后还是最优解，而 trPCA 的解经过正交变换之后还是最优解. 因此，如果基于欧氏距离处理聚类或分类问题，trPCA 的不同解得到的聚类或分类结果是不变的，但 dPCA 的不同解可能会使得投影后的距离发生改变，从而导致聚类或分类的结果不稳定.

**注 1**　当 $k=1$，即提取的主成分维数为一维时，trPCA 和 dPCA 方法是等价的，同为下述优化问题：

$$\max_{u\in\mathbf{R}^{d}}\frac{u^{\mathrm{T}}C_{xx}u}{u^{\mathrm{T}}C_{yy}u}.$$

**注 2**　当没有背景数据时，$C_{yy}=I_d$ 成立，trPCA 退化为 PCA.

**注 3**　根据文献[15]，$\mathrm{tr}(\hat{U}^{\mathrm{T}}C_{xx}\hat{U})/\mathrm{tr}C_{xx}$表示 $k$ 个主成分$\{z_i\}_{i=1}^{k}$的累计贡献率，因而 PCA 的最优解使得累计贡献率最高. 因极大化问题(4)的目标函数等价于使得$\dfrac{\mathrm{tr}(\hat{U}^{\mathrm{T}}C_{xx}\hat{U})/\mathrm{tr}C_{xx}}{\mathrm{tr}(\hat{U}^{\mathrm{T}}C_{yy}\hat{U})/\mathrm{tr}C_{yy}}$最大，可以解释为 trPCA 寻找使得目标数据累计贡献率与背景数据累计贡献率的比值最大的 $\hat{U}$.

### 3.2　背景方差奇异情形

为处理 $C_{yy}$ 奇异情形，我们将问题(4)等价地转化为下述问题：

$$\max_{U\in\mathbf{R}^{d\times k}}\frac{\mathrm{tr}(U^{\mathrm{T}}C_{xx}U)}{\mathrm{tr}(U^{\mathrm{T}}C_{t}U)},\ \text{s.t.}\quad U^{\mathrm{T}}U=I_k \tag{5}$$

其中 $C_t=C_{xx}+C_{yy}$. 问题(5)的目标函数值必然属于区间$[0,1]$，且最大值 1 对应于问题(4)的最大值 $\infty$. 我们对 $C_{xx}$ 的零空间$\{x\,|\,C_{xx}x=0\}$不感兴趣，这是因为 $x^{\mathrm{T}}C_{xx}x=0$，其零空间中不包含关于目标数据集的任何有用信息，把它们从问题(5)的约束空间中移除并不会牺牲特征提取准确度. 实际上，由于属于 $C_{xx}$零空间且不属于$C_{yy}$零空间的部分不可能是问题(5)的最优解，所以我们只需移除 $C_t$ 的零空间.

进一步，在很多实际场景中，涉及到越来越多的高维数据，样本协方差矩阵的条件数很大，甚至接近于奇异情形. 为了更好的特征提取效果，接近 $C_t$ 零空间的部分也要移除. 对 $C_t$ 进行特征分解可得

$$C_t=Q\Lambda Q^{\mathrm{T}} \tag{6}$$

其中 $\Lambda=[\lambda_1,\lambda_2,\cdots,\lambda_d]$，$\lambda_l\geqslant 0$，$l=1,2,\cdots,d$. 我们假设特征值已按由大到小顺序排列，$Q$ 为正交矩阵. 选取 $1\leqslant d'\leqslant d-1$ 使得

$$\frac{\lambda_{d'}+\lambda_{d'+1}+\cdots+\lambda_d}{\lambda_1+\lambda_2+\cdots+\lambda_d}\leqslant\varepsilon \tag{7}$$

其中阈值 $\varepsilon$ 为一个很小的标量. 取矩阵 $Q$ 的前 $d'$ 列组成矩阵 $W$. 解空间可以被限制在由 $W$ 的列空间张成的子空间里，即 $U=WV$，$V\in\mathbf{R}^{d'\times k}$. 从而(5)式定义的问题可以转化为如下问题，称为修正的 trPCA 方法：

$$\max_{V\in\mathbf{R}^{d'\times k}}\frac{\mathrm{tr}(V^{\mathrm{T}}C_{xx}^{u}V)}{\mathrm{tr}(V^{\mathrm{T}}C_{t}^{u}V)},\ \text{s.t. } V^{\mathrm{T}}V=I_k \tag{8}$$

其中 $C_{xx}^u = W^{\mathrm{T}} C_{xx} W$, $C_t^u = W^{\mathrm{T}} C_t W$. 因而 $C_t^u$ 恒为正定矩阵,不会产生奇异问题. 但 dPCA 无法处理 $C_{yy}$ 奇异情形. 此外,因 $C_{xx}^u$ 的维数小于 $C_{xx}$ 的维数,该算法降低了后续二分法以及特征分解的计算量. 基于上述分析,最终的投影矩阵 $\hat{U} = W\hat{V}$,其中 $\hat{V}$ 为问题(8)的最优解.

修正的 trPCA 算法的具体执行步骤如下.

**算法 3.3**　修正的迹比率主成分分析

1. 输入:目标数据 $\{x_i\}_{i=1}^n$、背景数据 $\{y_j\}_{j=1}^m$、提取的主成分维数 $k$ 和阈值 $\varepsilon$;

2. 将 $\{x_i\}_{i=1}^n$ 和 $\{y_j\}_{j=1}^m$ 中心化;

3. 计算经验协方差矩阵 $C_{xx}$ 和 $C_{yy}$;

4. 利用(6)式进行特征分解;

5. 利用(7)去除方差阵接近零空间的部分;

6. 利用二分法[10]求解问题(8);

7. 输出:子空间 $\hat{U} \in \mathbf{R}^{d\times k}$.

**注 4**　当 $C_{yy}$ 奇异时,由问题(8)可得到和定理 3.1 类似的结论.

### 3.3　多背景数据集情形

在某些情形下,我们更倾向于寻找某一个数据集相对于多个背景数据集所特有的判别信息. 假设除了 $\{x_i\}_{i=1}^n$,还给定 $M \geqslant 2$ 个背景数据集 $\{y_j^s\}_{j=1}^{m_s}$, $s = 1, \ldots, M$, 背景数据集中包含和 $\{x_i\}$ 相同的潜在背景子空间向量,并假设所有数据都已被中心化且维数为 $d$. 令

$$C_{xx} := \frac{1}{n}\sum_{i=1}^{n} x_i x_i^{\mathrm{T}}, C_{yy}^s := \frac{1}{m_s}\sum_{j=1}^{m_s} y_j^s (y_j^s)^T$$

为相应的样本协方差矩阵. 我们的目标是挖掘出能明显表示目标数据,同时又不属于任何背景数据的潜在子空间向量.

基于 dPCA 算法,文献[5]提出了多背景数据判别主成分分析(MdPCA),该方法求解以下问题:

$$\max_{U\in\mathbf{R}^{d\times k}} \mathrm{tr}\Big[\Big(\sum_{s=1}^{M}\omega_s U^{\mathrm{T}} C_{yy}^s U\Big)^{-1} U^{\mathrm{T}} C_{xx} U\Big],$$

其中 $\{\omega_s \geqslant 0\}_{s=1}^M$,满足 $\sum_{s=1}^{M}\omega_s = 1$. 同样,由于 dPCA 算法的局限性,MdPCA 算法也有待改进.

基于前面对单一背景数据的分析,我们应该寻找使得目标方差极大,且所有的背景方差极小的方向. 形式上,我们求解以下极大化问题,称为多背景数据迹比率主成分分析(MtrPCA)算法:

$$\max_{U\in\mathbf{R}^{d\times k}} \frac{\mathrm{tr}(U^{\mathrm{T}} C_{xx} U)}{\mathrm{tr}\Big(\sum_{s=1}^{M}\omega_s U^{\mathrm{T}} C_{yy}^s U\Big)}, \ \text{s.t. } U^{\mathrm{T}}U = I_k \tag{9}$$

其中 $\{\omega_s \geqslant 0\}_{s=1}^M$,满足 $\sum_{s=1}^{M}\omega_s = 1$ 为 $M$ 个背景方差各自的权重. 提取后的主成分称为多背景数据迹比率主成分(MtrPCs)算法. 如果令 $C_{yy} := \sum_{s=1}^{M}\omega_s C_{yy}^s$,则问题(9)退化为问题(4). 接下来的求解方法类似于 trPCA 算法. 此外对于方差阵接近奇异的情形也有相应的修正算法.

**注 5**　对于参数 $\omega_s$ 的选取有两种方法:

(i) 用谱聚类[6]的方法选出少数几个可以产生最具代表性子空间的 $\omega_s$;

(ii) 将问题(9)中的 $\omega_s$ 与 $U$ 联合进行优化.

## 4　仿　真

为了验证本文所提算法在进行判别分析时的性能,本节提供了 4 个数值算例,与已有算法进行对比.

**例 4.1**　首先,我们利用真实的老鼠蛋白质表达数据[16]来验证 trPCA 对真实数据的特征提取能力. 目标数据 $\{x_i \in \mathbf{R}^{77}\}_{i=1}^{270}$ 包含 270 个数据,每个数据由 77 维蛋白质表达数据组成. 目标数据都来自于患有唐氏综合征的老鼠,且前 135 个数据 $\{x_i\}_{i=1}^{135}$ 记录的是 135 个接受了药物治疗的老鼠的观测,后 135 个数据 $\{x_i\}_{i=136}^{270}$ 记录的是作为对照组注射盐水的老鼠的观测. 背景数据 $\{y_j \in \mathbf{R}^{77}\}_{j=1}^{135}$ 来自于对 135 只健康老鼠的观测,它们和患病老鼠的性别、年龄等信息呈现类似分布,只是没患有唐氏综合征. 对于部分缺失数据,我们采用了文献[16]中的做法,将其补为同类别老鼠蛋白质观测数据的平均值.

我们对数据集 $\{\{x_i\},\{y_j\}\}$ 应用了 trPCA、dPCA 与 cPCA 算法,其中应用 trPCA 算法时(7)式中的 $\varepsilon$ 取成 0.001, cPCA 算法的参数 $\alpha$ 选为 1, 10 和 100. 单独对 $\{x_i\}$ 进行 PCA 算法处理. 为了便于展示,所有提取的主成分维数都为 2,并且前两个主成分即为图 1 的横、纵坐标. 如图 1 所示,圆圈表示接受药物治疗的老鼠,叉号表示没有接受治疗的老鼠,trPCA 可以较好地区分两个不同的子类,而 PCA 和 dPCA 算法不能很好的区分两个子类. cPCA 算法对 $\alpha$ 的不同选值较为敏感,特征提取效果不稳定.

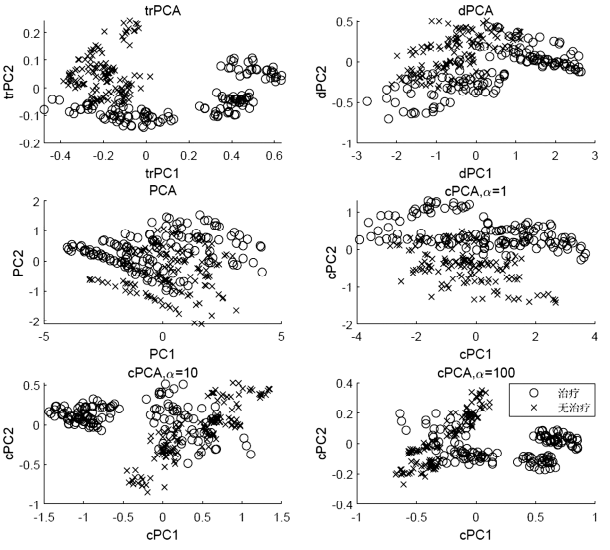


图 1　寻找老鼠蛋白质表达数据中的子类

Fig. 1　Discovering subgroups in mice protein expression data

**表 1　基于老鼠蛋白质表达数据特征提取的聚类误差**

Tab. 1　Clustering error based on feature extraction of mice protein expression data

| $k$ | trPCA | dPCA | PCA | cPCA ($\alpha=1$) | cPCA ($\alpha=10$) | cPCA ($\alpha=100$) |
|---|---|---|---|---|---|---|
| 1 | 0.222 2 | 0.292 5 | 0.414 8 | 0.374 1 | 0.266 7 | 0.222 2 |
| 2 | 0.222 2 | 0.296 2 | 0.396 2 | 0.374 1 | 0.248 1 | 0.222 2 |
| 3 | 0.218 5 | 0.296 2 | 0.396 2 | 0.374 1 | 0.266 7 | 0.222 2 |
| 4 | 0.218 5 | 0.307 4 | 0.396 3 | 0.374 1 | 0.248 1 | 0.222 2 |
| 5 | 0.214 8 | 0.359 2 | 0.396 2 | 0.374 1 | 0.251 9 | 0.222 2 |
| 10 | 0.222 2 | 0.362 9 | 0.400 0 | 0.374 1 | 0.266 7 | 0.222 2 |

为了进一步衡量不同算法的表现，我们先用这些算法提取出目标数据的 $k$ 维主成分，再对提取后的低维表示通过 K-均值聚类分成两类，最后比较这几种算法的聚类误差. 聚类误差定义为错误聚类的数据个数与目标数据的总个数的比值. 如表 1 所示，trPCA 算法的聚类误差要小于 dPCA 和 PCA 算法，cPCA 算法的聚类误差只有当选取合适的 $\alpha$ 时才会低一些，但 $\alpha$ 的选取会产生很高的计算复杂度. 因而在后面的实验中我们主要和 dPCA 算法进行比较.

**例 4.2**　当背景方差矩阵 $C_{yy}$ 奇异时，dPCA 算法无法处理，而 trPCA 算法可以通过解修正的 trPCA 算法来避免 $C_{yy}$ 奇异所带来的困扰. 在第二个实验中，为了检验 trPCA 算法处理奇异情形的表现，我们考虑由两个目标子集构成的目标数据 $\{x_i\}$. 目标子类 1 包含 120 个 280 维向量，前 200 维 $N(0,1)$，后 80 维来自于 $N(0,10)$. 目标子类 2 也包含 120 个 280 维向量，其中前 200 维由 $N(6,1)$生成，后 80 维由 $N(0,10)$生成. 背景数据集 $\{y_j\}$包含 240 个 280 维向量，前 200 维来自于 $N(0,3)$，后 80 维来自于 $N(0,10)$，详见表 2. 此时 $C_{yy}$ 显然是不可逆的.

**表 2　背景方差奇异情形下的数据生成分布假设汇总**

Tab. 2　Summary of the distributional assumptions used to generate the data used in background covariance singular settings

| | 子类 1 | 子类 2 | 背景数据 |
|---|---|---|---|
| 1～200 维 | $N(0,1)$ | $N(6,1)$ | $N(0,3)$ |
| 201～280 维 | $N(0,10)$ | $N(0,10)$ | $N(0,10)$ |

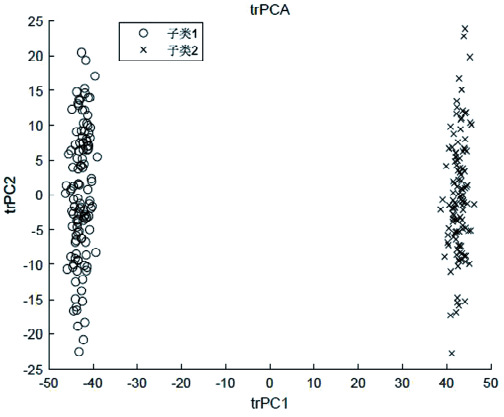


图 2　背景方差奇异时的特征提取

Fig. 2　The feature extraction with singular background covariance

求解问题(8)，提取前两个主成分 trPC1 和 trPC2 可以得到如图 2 的结果，其中的圆圈表示子类 1 的数据，叉号表示子类 2 的数据. 由图可见，trPCA 算法可以很好的处理背景方差奇异的情形，并且能有效提取出目标子类相对背景数据的判别信息. dPCA 算法则无法处理此类问题.

**例 4.3**　接下来的数值模拟检验 trPCA 算法在处理图像识别问题时的表现. 通过叠加 MNIST 和 CIFAR-10[17]数据库的数据，我们得到半合成数据. 目标数据 $\{x_i \in \mathbf{R}^{784}\}_{i=1}^{2\,000}$ 由 2 000 个 28×28 的手写数字 1 和 2(各 1 000 个)叠加上 2 000 个来自于 CIFAR-10 数据库的青蛙图片生成，且每个数据都已经过中心化处理. 其中，青蛙图片是通过把未加工的 32×32 青蛙图片裁掉四周，截取中间的 28×28 图像，并转化成灰度图. 背景数据 $\{y_j \in \mathbf{R}^{784}\}_{j=1}^{3\,000}$ 由 3 000 个裁剪过后的青蛙图片组成.

我们对数据集$\{\{x_i\},\{y_j\}\}$应用了 trPCA 算法和 dPCA 算法,并单独对$\{x_i\}$应用 PCA 算法. 提取的主成分维数为 2. 如图 3 所示,点表示数字 1,叉号表示数字 2,trPCA 的判别效果要优于其它两种算法.

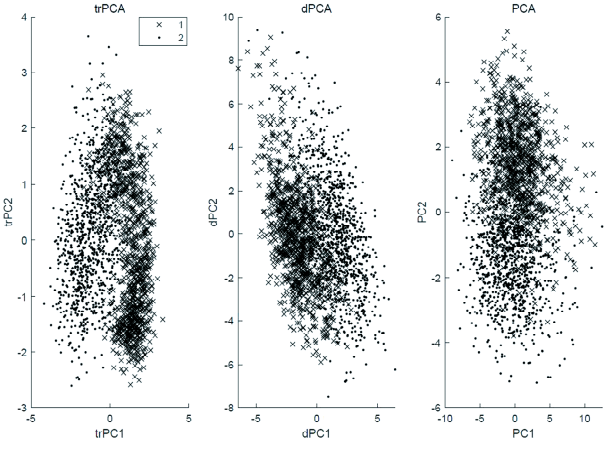


图 3　单背景叠加手写数字 1 和 2 的特征提取

Fig. 3　The feature extraction of single background superimposed handwritten digits 1 and 2

**表 3　基于单背景叠加手写数字特征提取的聚类误差**

Tab. 3　Clustering error based on feature extraction of single background superimposed handwritten digits

| $k$ | trPCA | dPCA | PCA |
|---|---|---|---|
| 1 | 0.113 5 | 0.158 0 | 0.422 5 |
| 2 | 0.108 0 | 0.365 0 | 0.422 5 |
| 3 | 0.121 5 | 0.395 0 | 0.　424 0 |
| 4 | 0.102 0 | 0.403 5 | 0.424 0 |
| 5 | 0.101 5 | 0.403 5 | 0.424 0 |
| 10 | 0.104 5 | 0.417 5 | 0.422 0 |

接下来,在提取出目标数据的 $k$ 维表示后,我们进行 K-均值聚类,表 3 给出这几种算法的聚类误差. 可见 trPCA 算法的聚类效果远好于其它两种算法.

为了展示 trPCA 算法在处理高维数据时的高效性,下表给出了这几种算法在 MATLAB 中处理上述问题时 $k$ 从 1 取到 10 的运行总时间. 可见 trPCA 算法的运行时间小于其它算法,体现了 trPCA 算法相对于其它几种算法的高效性.

**表 4　几种算法的运行时间比较**

Tab. 4　Comparison of running time of several algorithms

| | trPCA | dPCA | PCA |
|---|---|---|---|
| 时间/s | 33.59 | 38.61 | 35.26 |

**例 4.4**　为了检验 MtrPCA 算法在处理多背景数据集时的功效,类似于上一个实验,我们叠加 MNIST 和 CIFAR-10 数据库的数据. 具体而言,目标数据$\{x_i\in\mathbf{R}^{784}\}_{j=1}^{400}$由 400 个 28×28 的手写数字 1 和 2(各 200 个)叠加上 400 个来自于 CIFAR-10 数据库的船图片生成,并且每个数据都已被中心化. 背景数据$\{y_j^1\in\mathbf{R}^{784}\}_{j=1}^{200}$的前 392 维和$\{y_j^2\in\mathbf{R}^{784}\}_{j=1}^{200}$的后 392 维分别对应于 200 个裁剪成 28×28 后的船图片的前 392 维和后 392 维. 所有背景数据的剩余维数的数据的都设为 0. 对上述多背景数据进行 MtrPCA 算法和 MdPCA 算法[5],权重系数都设为 $\omega_1=\omega_2=0.5$,单独对目标数据进行 PCA 算法. 对比结果如下图所示,叉号和圆圈分别表示数字 1 和数字 2 的数据点的前两个主成分. MtrPCA 算法可以很好的区分两个子类,但 MdPCA 算法和 PCA 算法效果较差,无法将两个子类分开.

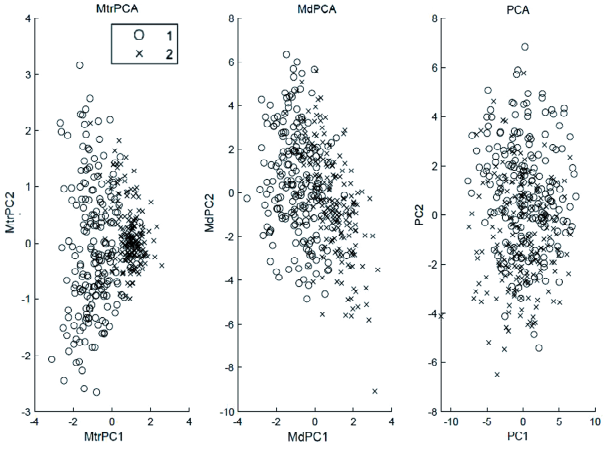


图 4　多背景叠加手写数字 1 和 2 的特征提取

Fig. 4　The feature extraction of multiple background superimposed handwritten digits 1 and 2

接下来进行 K-均值聚类,当提取的特征维数不同时,相应的聚类误差如下表所示. 可以看到 MtrPCA 算法的聚类误差远小于 MdPCA 算法和 PCA 算法.

**表 5　基于多背景叠加手写数字数据特征提取的聚类误差**

Tab. 5　Clustering error based on feature extraction of multiple background superimposed handwritten digits

| $k$ | MtrPCA | MdPCA | PCA |
|---|---|---|---|
| 1 | 0.102 5 | 0.167 5 | 0.440 0 |
| 2 | 0.097 5 | 0.360 0 | 0.452 5 |
| 3 | 0.092 5 | 0.395 0 | 0.455 0 |
| 4 | 0.092 5 | 0.395 0 | 0.455 0 |
| 5 | 0.085 0 | 0.410 0 | 0.455 0 |
| 10 | 0.085 0 | 0.465 0 | 0.442 5 |

## 5 结　论

在各种实际场景中，我们更希望能找出某个数据集相对于其它数据集所具有的独特判别信息. 本文提出了一种新方法，称为 trPCA 算法，对多个数据集进行判别分析. 该算法可推广至多背景数据模型 MtrPCA 算法. 与 dPCA 算法相比，trPCA 算法可以处理背景方差矩阵奇异的情形，特征提取效果以及后续处理分类或聚类问题也都优于 dPCA 算法. 作为 cPCA 算法的改进，trPCA 算法不需要额外的超参数且有高效的算法进行求解，计算复杂度低. 最后，多个数值模拟表明本文所提算法相较于已有算法具有更好的性能.

**参考文献：**


[1] 李培江，李碧娟，南伟，等. 油用与食用向日葵籽形态及主成分差异辨析 [J]. 四川大学学报：自然科学版，2017，54：203.

[2] 边喜丽，杨小林，李永霞，等. 藏东南巨柏根系结构特征与环境因子研究 [J]. 四川大学学报：自然科学版，2018，55：848.

[3] Pearson K. LIII. On lines and planes of closest fit to systems of points in space [J]. Phil Mag J Sci，1901，2：559.

[4] Abid A，Zhang M J，Bagaria V K，*et al*. Exploring patterns enriched in a dataset with contrastive principal component analysis [J]. Nat Commun，2018，9：2134.

[5] Chen J，Wang G，Giannakis G B. Nonlinear dimensionality reduction for discriminative analytics of multiple datasets [J]. IEEE T Signal Process，2018，67：740.

[6] Ng A Y，Jordan M I，Weiss Y. On spectral clustering：analysis and an algorithm [C]//Advances in neural information processing systems. Cambridge：MIT，2002：849.

[7] Yang Q，Bassyouni A，Butler C R，*et al*. Ligand biological activity predicted by cleaning positive and negative chemical correlations [J]. P Natl A Sci India A，2019，116：3373.

[8] Chen J，Wang G，Shen Y，*et al*. Canonical correlation analysis of datasets with a common source graph [J]. IEEE T Signal Process，2018，66：4398.

[9] Fukunaga K. Introduction to statistical pattern recognition [M]. New York：Elsevier，2013.

[10] Guo Y F，Li S J，Yang J Y，*et al*. A generalized Foley - Sammon transform based on generalized fisher discriminant criterion and its application to face recognition [J]. Pattern Recogn Lett，2003，24：147.

[11] Jia Y，Nie F，Zhang C. Trace ratio problem revisited [J]. IEEE T Neur Net Lear，2009，20：729.

[12] Wang H，Yan S，Xu D，*et al*. Trace ratio *vs*. ratio trace for dimensionality reduction [C]//2007 IEEE conference on computer vision and pattern recognition. New York：IEEE，2007：1.

[13] Belhumeur P N，Hespanha J P，Kriegman D J. Eigenfaces *vs*. fisherfaces：recognition using class specific linear projection [J]. IEEE T Pattern Anal，1997：711.

[14] Yan S，Xu D，Zhang B，*et al*. Graph embedding：a general framwork for dimensionality reduction [C]//2005 IEEE computer society conference on computer vision and pattern recognition. New York：IEEE，2005：830.

[15] 高惠璇. 应用多元统计分析 [M]. 北京：北京大学出版社，2005.

[16] Higuera C，Gardiner K J，Cios K J. Self-organizing feature maps identify proteins critica to learning in a mouse model of down syndrome [J]. Plos One，2015，10：e0129126.

[17] Krizhevsky A，Hinton G. Learning multiple layers of features from tiny images [R]. Toronto：University of Toronto，2009.



**引用本文格式：**

中　文：赵小彤，李智，宋恩彬. 基于迹比率的多数据集判别分析维数压缩[J]. 四川大学学报：自然科学版，2021，58：011002.

英　文：Zhao X T，Li Zhi，Song E B. Trace ratio-based dimensionality reduction for discriminative analysis of multiple datasets [J]. J Sichuan Univ：Nat Sci Ed，2021，58：011002.