中国理论数学前沿
Frontiers of Chinese Pure Mathematics
2023 年第 01 卷第 02 期
Vol. 01, No. 02, 2023

SCIENCE FOOTPRINT PRESS

研究性论文
http://www.scifootprint.com/FCPM/index.html

基于 l_q 正则项的稀疏线性判别分析

陈　静，高彩霞*

（内蒙古大学数学科学学院，呼和浩特 010021）

摘要：线性判别分析在特征提取和数据降维及分类上具有重要的作用，随着科技的进步，需要处理的数据越来越庞大，然而在高维情形下，线性判别分析面对着两个问题—投影后的数据解释性不足，因为它们均涉及到所有 p 个特征，是所有特征的线性组合以及类内协方差矩阵的奇异性问题。线性判别分析存在三个不同的论点：多元高斯模型、Fisher 判别问题和最优评分问题。针对这两个问题，本文建立了一种求解第 k 个判别成分的模型，该模型首先对线性判别分析中的 Fisher 判别问题原始模型做出变换，利用类内方差的对角估计矩阵代替原始类内协方差矩阵，克服了矩阵奇异的问题，同时将其投影到正交投影空间上，以便去掉其正交约束，随后加入了 l_q 范数正则项，增强其解释性，实现降维和分类的目的。最后给出了求解该模型的迭代算法及收敛性分析，证明了由该算法产生的序列是下降收敛的，且对任意初始值均收敛到问题的局部最小值。

关键词：线性判别分析；稀疏优化；l_q 范数；投影

DOI：10.48014/fcpm.20230529001

引用格式：陈静，高彩霞．基于 l_q 正则项的稀疏线性判别分析[J]．中国理论数学前沿，2023，1(2)：31-38.

0　引言

模式分类是模式识别的核心研究内容，它可以分为有监督学习和无监督学习。有监督学习需要已知每个数据点所属的类别或标签作为输入，并通过训练来建立一个分类模型，使其能够正确地预测新数据点的类别或标签。无监督学习则不需要提供标签信息，而是寻找数据中的内在结构并将数据分组到不同的簇中。分类算法是模式分类的核心部分之一，常见的分类算法包括决策树、朴素贝叶斯、支持向量机、神经网络等。

近年来受到广泛关注的特征选择和提取在模式分类中发挥着重要作用。然而，图像原始数据包含大量冗余特征和噪声，给图像识别和图像分析带来了困难[1]。在这种情况下，对于分类任务，如何选择和提取整个练习中最重要的特征的不同类别是最重要的，也是最难的。事实证明，特征选择和提取是机器学习和模式分类领域的有效工具。它们可以降低复杂性，提高效率并增强分类性能[2]。

在过去的几十年中，人们提出了各种特征提取方法。在这个分支中，主成分分析（PCA）是最著名的方法之一，其主要思想是尝试学习一个可以保留原始数据主要能量的投影[3]。局部保留投影（LPP），稀疏保留投影（SPP）和邻域保留嵌入（NPE），他们学习的投影也是特征提取方法，这些方法考虑了原始的局部流形几何数据并尝试在投影空间中保留局部信息。随后，分类器，例如 K 最近邻（KNN）和支持向量机（SVM），通常用于分类。上述方法虽然没有使用数据的标签信息，各有优势，但在一定程度上这些算法的分类性能还不够好。

实际上，我们希望对带有类别标签的数据（有

* 通讯作者　Corresponding author：高彩霞，2954993678@qq.com
收稿日期：2023-05-29；　录用日期：2023-06-21；　发表日期：2023-09-28

监督的)可以实现降维,降维后可以更好地区分每个类别。这时,出现了另一种经典的特征提取算法—线性判别分析(LDA)。然而在高维情形下,从LDA得到的分类规则缺乏解释性,因此,为获得稀疏的判别变量,本文将在LDA中引入 l_q 范数正则项。

1 预备知识

1.1 符号与定义

$\alpha=(a_1,a_2,\cdots,a_p)^{\mathrm{T}}\in\mathbb{R}^p$,其中 $(\cdot)^{\mathrm{T}}$ 表示向量或矩阵的转置,$\mathbb{R}^p$ 表示 p 维实数向量的全体。$\boldsymbol{X}\in\mathbb{R}^{n\times p}$ 是一个数据矩阵,其中 $\mathbb{R}^{n\times p}$ 表示 $n\times p$ 维实数矩阵的全体。x_{ij} 表示矩阵的第 i 行第 j 列元素。$\mathrm{Tr}(\boldsymbol{X})$ 表示矩阵 $\boldsymbol{X}$ 的迹,即主对角线上各元素之和,$\mathrm{rank}(\boldsymbol{X})$ 表示矩阵 $\boldsymbol{X}$ 的秩,$\Sigma=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}/n$ 表示矩阵 $\boldsymbol{X}$ 的协方差矩阵。广义实数集可表示为 $\overline{\mathbb{R}}=\mathbb{R}\cup\{\pm\infty\}$。

1.2 文献综述

线性判别分析(Linear discriminant analysis)因其简单、稳健和预测准确性而成为许多应用中监督分类的首选工具[4]。LDA还提供数据在最具辨别力的方向上的低维投影,这对于数据解释很有用。LDA分类器存在三个不同的论点:多元高斯模型、Fisher判别问题和最优评分问题。经典LDA通常在简单和低维数据集上具有优异的性能。但已知它存在两个问题:

(1)当预测变量 p 的数量大于观察值 n 时。在这种情况下,不能直接应用LDA,因为特征的类内协方差矩阵是奇异的[5]。

(2)当 p 较大时,从LDA得到的分类规则很难解释,因为它涉及所有 p 特征,是所有特征的线性组合。

在某些情况下 $n\leqslant p$,人们可能希望有一个执行特征选择的分类器。这种稀疏分类器确保更容易的模型解释,并可以减少训练数据的过度拟合。

Fisher LDA通过最大化类间方差和类内方差之间的样本比率来寻找一组判别向量。为解决特征的类内协方差矩阵奇异这一问题,Krzanowski等[6]考虑了类内协方差矩阵的其他正定估计。Guo、Hastie和Tibshirani等[7]在萎缩质心正则化判别分析(RDA)中,估计协方差矩阵时结合了脊型罚和软阈值。

为获得稀疏判别向量,受Zou等[8]提出的稀疏PCA方法启发,Qiao等[9]提出了针对高维低样本数据的稀疏线性判别分析。首先通过将广义特征值问题转换为回归类型问题,将判别向量与回归系数向量相关联,然后应用具有 l_1 惩罚的惩罚最小二乘法来解决问题。2011年Witten等[10]在其文献中同样将 l_1 惩罚应用于Fisher判别问题,解决该问题是具有挑战性的,因为它不是凸问题,因此必须应用专门的技术,如文中所提出的最小化最大化方法。同年,Shao等[11]提出了基于高维数据阈值的稀疏线性判别分析,证明了其在未知参数的某些稀疏条件下渐近最优。文章中还表明,当样本数 $n\to\infty$时,Fisher线性判别是渐近最优的。但是当变量维数 p 远大于 n 时,Fisher判别准则的错判概率接近 $\frac{1}{2}$,也就是说此时线性判别分析相当于随机猜测。

作为LDA的一种等价形式,开发了最优评分,这在Hastie、Buja和Tibshirani的文章中进行了详细讨论[12]。它是通过一系列评分,将分数分配给类别(组、类别),将分类变量转换为定量变量,从而将分类问题重铸为回归问题。对于二值分类,Mai和Zou[21]建立了Fisher判别问题和最优评分问题之间的等价性,并声明它们都按照缩放正则化参数的贝叶斯规则进行求解。最优评分标准的形式为:

$$\begin{aligned}&\min_{\beta_k,\theta_k}\quad \|\mathrm{Y}\theta_{\mathrm{k}}-X\beta_{\mathrm{k}}\|\\&\text{s.t}\quad \frac{1}{n}\theta_k^{\mathrm{T}}\mathrm{Y}^{\mathrm{T}}\mathrm{Y}\theta_k=1,\\&\qquad\theta_k^{\mathrm{T}}Y^{\mathrm{T}}Y\theta_l=0.\ (1\leqslant l\leqslant k-1)\end{aligned}$$

式中,$\boldsymbol{X}$ 是一个 $n\times p$ 数据矩阵,在行上具有观察值,在列上具有特征,$\boldsymbol{Y}$ 表示 k 类伪变量的 $n\times k$ 矩阵,Y_{ik} 是第 i 个观测值是否属于第 k 类的指标变量,如果第 i 个观测值属于第 k 类,则由 $Y_{ij}=1$ 定义,否则有 $Y_{ij}=0$。由于 $\boldsymbol{X}$ 的列居中,平均值为零[13]。一些研究人员提出通过对 β_k 施加惩罚来惩罚最优得分标准。

2 模型

2.1 模型介绍

下面介绍 Fisher 线性判别分析(FLDA 或 LDA)的相关定义和模型,FLDA 目的是寻找一个使得类间平方和与类内平方和的比率最大化的线性函数 $a^{\mathrm{T}}x$,形式上,假设有 k 个类别,并设 x_{ij},$j=1,\cdots,n_i$,$i=1,\cdots,k$,表示第 i 类观测值的向量。x_i 表示第 i 类的平均值,$n=n_1+\cdots+n_k$,令 $\boldsymbol{X}$ 是一个 $n\times p$ 数据矩阵,且 $y=\boldsymbol{X}a$,则 FLDA 求解:

$$\max_a \frac{\sum_{i=1}^k n_i(\bar{y}_i-\bar{y})^2}{\sum_{i=1}^k\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2}.$$

式中,$\bar{y}_i$ 表示 y 的第 i 个子向量 y_i 的均值,对应于第 i 个类别。用 $\boldsymbol{X}a$ 代替 y,我们可以将类内平方和重写为:

$$\begin{aligned}&\sum_{i=1}^k\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2\\&=a^{\mathrm{T}}\sum_{i=1}^k\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)(x_{ij}-\bar{x}_i)^{\mathrm{T}}a\\&\overset{\mathrm{def}}{=\!=}a^{\mathrm{T}}\sum\nolimits_w a.\end{aligned}$$

类间平方和为:

$$\begin{aligned}&\sum_{i=1}^k n_i(\bar{y}_i-\bar{y})^2\\&=\sum_{i=1}^k n_i\{a^{\mathrm{T}}(\bar{x}_i-\bar{x})\}^2\\&=a^{\mathrm{T}}\sum_{i=1}^k n_i(\bar{x}_i-\bar{x})(\bar{x}_i-\bar{x})^{\mathrm{T}}a\\&\overset{\mathrm{def}}{=\!=}a^{\mathrm{T}}\sum\nolimits_b a.\end{aligned}$$

因此,比率可以由下式给出:

$$\frac{a^{\mathrm{T}}\sum_b a}{a^{\mathrm{T}}\sum_w a}.$$

如果 a_1 是使比率最大化的向量,则可以找到在 $\sum_w$ 中与 a_1 正交的下一个方向 a_2,使得比率最大化,并且可以类似地依次计算附加方向。投影方向 a_i 通常称为判别坐标,线性函数 $a_i^{\mathrm{T}}x$ 称为 Fisher 线性判别函数。

2.2 模型发展

下面介绍稀疏 LDA 的模型发展,LDA 的主要思想是投影,该算法的优点是缩短了同类型样本的投影距离,同时也可以增加不同类型的样本的距离。但是 LDA 也有它的设计缺陷。它的大部分投影系数不为零,新的特征样本是所有样本特征的线性组合,导致投影后的矩阵对特征缺乏很好的解释性,体现在以下几个方面。首先,LDA 对矩阵特征没有很好的解释,无法在大量冗余数据中选择最合适的函数。其次,鲁棒性较差,LDA 选择前 k 个最小特征值对应的 k 个特征向量作为特征提取投影时,受数据影响较大,LDA 数据分类变化较大。这导致对于不同大小的数据集,LDA 的分类准确率存在很大差异。针对这些问题,众多研究者提供了解决方法。例如,一种获得稀疏判别变量的方法是利用 l_1 惩罚,其优化问题为:

$$\begin{aligned}\max_{\beta_k}\quad&\beta_k^{\mathrm{T}}\sum\nolimits_b\beta_k-r\|\beta_k\|_1\\\text{s. t.}\quad&\beta_k^{\mathrm{T}}(\sum\nolimits_w+\Omega)\beta_k=1,\\&\beta_k^{\mathrm{T}}(\sum\nolimits_w+\Omega)\beta_l=0.(\forall\, l<k)\end{aligned}$$

Witten 和 Tibshirani 在文献[10]中采用了这种方法,并采取最小化—最大化方法求解该模型。

2022 年束磊等[14]将 l_0 惩罚用于原始模型上,提出了基于 l_0 约束的稀疏线性判别分析模型如下:

$$\begin{aligned}&\max_{\beta_k}\{\beta_k^{\mathrm{T}}\widehat{\sum}_b^k\beta_k\}\\&\text{s. t.}\quad\|\beta_k\|_0\leqslant s,\\&\qquad\beta_k^{T}\widetilde{\sum}_w\beta_k=1.\end{aligned}$$

式中,$\widetilde{\sum}_w$ 是类内协方差矩阵的对角阵估计,$\widehat{\sum}_b^k$ 是利用正交投影思想转化后的类间协方差矩阵,s 是事先给定的稀疏度。文中给出了一种稀疏 LDA 算法进行求解,在实验效果上说明了其收敛性,但并未证明。

Line Vlemmensen 等[5]转而应用 l_1 惩罚到 LDA 的最佳得分公式中。文中的 SDA 方法是按顺序定义的。第 k 个 SDA 解对(θ_k,β_k)求解问题:

$$\begin{aligned}&\min_{\beta_k,\theta_k}\quad\|\mathrm{Y}\theta_k-\mathrm{X}\beta_k\|+r\beta_k^{\mathrm{T}}\Omega\beta_k+\lambda\|\beta_k\|_1\\&\text{s. t.}\quad\frac{1}{n}\theta_k^{\mathrm{T}}Y^{\mathrm{T}}Y\theta_k=1,\\&\theta_k^{\mathrm{T}}Y^{\mathrm{T}}Y\theta_l=0\quad(1<l\leqslant k-1)\end{aligned}$$

式中,λ 为非负调谐参数,$\boldsymbol{\Omega}$ 为正定惩罚矩阵。如果对于 $r>0$,$\boldsymbol{\Omega}=rI$ 那么这是一个弹性净罚。如果 r 足够大,则生成的判别向量将是稀疏的。如果 $r=0$,

那么这就简化为 Hastie 等[5]提出的惩罚判别分析。上式可以用一种简单的迭代方式进行优化：我们对 β_k 保持固定 θ_k 进行优化，我们对 θ_k 保持 β_k 固定进行优化。

在文献[15]中，Le Thi 等将 l_0 正则化项添加到最优评分的目标函数以赋予判别向量稀疏性，提出 l_0 稀疏最优评分问题，其公式如下：

$$\min_{\beta_k,\theta_k} \quad \| Y\theta_k - X\beta_k \| + \lambda_1 \| \beta_k \|_2^2 + \lambda_2 \| \beta_k \|_0$$

$$\text{s. t.} \quad \frac{1}{n}\theta_k^{\mathrm{T}} Y^{\mathrm{T}} Y \theta_k = 1,$$

$$\theta_k^{\mathrm{T}} Y^{\mathrm{T}} Y \theta_l = 0 \quad (1 < l \leqslant k-1)$$

文中提出了一种基于 DC(凸函数差分)规划方法的新 BCD 方法。为了处理 l_0 正则化项，使用两个非凸连续函数来近似 l_0 范数。然后通过进行 DC 分解，利用直流算法解决最小化 l_0 的子问题。然而，在 DC 方案中，需要解决一系列非光滑子问题。这将增加新提出的坐标下降法的计算负担。

在线性判别分析的另一模式——最优评分问题中，李国权等[16]在 2021 年提出了加入 l_{q^-} 正则项的稀疏最优评分问题，并用迭代算法进行求解。受此系列文章启发，我们提出以下基于 l_{q^-} 正则项的稀疏 Fisher 线性判别模型。

2.3 加入 l_{q^-} 正则项的模型

Fisher LDA 寻找线性函数 $\beta^{\mathrm{T}} x$，使得类间平方和与类内平方和的比率最大化，形式上，考虑一个简单的多元高斯模型，即第 k 类的分布为 $N(\mu_k, \sum_w)$，$\mu_k \in \mathbb{R}^p$ 是均值向量，$\sum_w$ 是 $p \times p$ 的类内方差矩阵，假设有 K 个类别并设 $x_{ij}, j = 1, \cdots, n_i$ 是第 i 类观测值的向量，$i = 1, \cdots, k$，令 $n = n_1 + \cdots + n_k$，$\bar{x}_i$ 表示第 i 类的平均值。则 Fisher 判别问题可表示为：

$$\beta_k = \underset{\beta_k}{\operatorname{argmin}} \frac{\beta_k^{\mathrm{T}} \sum_w \beta_k}{\beta_k^{\mathrm{T}} \sum_b \beta_k}$$

$$\text{s. t. } \beta_k^{\mathrm{T}} \sum_w \beta_l = 0. \ (\forall\, l \leqslant k)$$

μ_k 的估计为 $\| C_k \|^{-1} \sum_{i \in C_k} x_i$，$C_k$ 表示第 k 类的指标集。π_k 是第 k 类的先验概率，$\sum_w$ 的估计为 $n^{-1} \sum_{k=1}^{K} \sum_{i \in C_k} (x_i - \mu_k)(x_i - \mu_k)^{\mathrm{T}}$，$\sum_b$ 的估计为 $n^{-1} \sum_{k=1}^{K} \pi_k \mu_k \mu_k \mu_k^{\mathrm{T}}$。

标准的组内方差协差阵估计量为：

$$\widehat{\sum_w} = n^{-1} \sum_{k=1}^{K} \sum_{i \in C_k} (x_i - \widehat{\mu_k})(x_i - \widehat{\mu_k})^{\mathrm{T}}.$$

其中，$\widehat{\mu_k}$ 是第 k 类的样本均值向量。

标准的组间方差 $\widehat{\sum_b}$ 估计量为：

$$\widehat{\sum_b} = n^{-1} x^{\mathrm{T}} x - \widehat{\mu_k} = n^{-1} \sum_{k=1}^{K} n_k \widehat{\mu_k} \widehat{\mu_k}^{\mathrm{T}}.$$

为解决矩阵奇异问题，一种简单方法是使用广义逆矩阵，例如 Duintjer Tebbens[17]在 2007 年提到的 Moore-Penrose 伪逆矩阵。尽管它很简单，但这种方法可能性能不佳，因为同年，Guo 等[18]已表明广义逆由于缺乏观测值而非常不稳定。Krzanowski 等[6]在其文献中考虑修改原问题，转而寻找一个单位向量 β，最大化 $\beta^{\mathrm{T}} \widehat{\sum_b} \beta$，使得 $\beta^{\mathrm{T}} \widehat{\sum_w} \beta = 1$。还有一种常用方法是在 Fisher 判别问题中使用类内协方差矩阵的正则化估计。例如：

$$\max_{\beta_k} \beta^{\mathrm{T}} \sum_b \beta_k$$

$$\text{s. t} \quad \beta_k^{\mathrm{T}} (\sum_w + \Omega) \beta_k = 1,$$

$$\beta_k^{\mathrm{T}} (\sum_w + \Omega) \beta_l = 1. \ (\forall\, l \leqslant k)$$

Hastie、Buja 和 Tibshirani 在文章中采用了这种方法，$\sum_w + \Omega$ 是正定的，因此，即使 $p \gg n$，问题中的判别向量也可以计算，此外，合适的 Ω 可以得出平滑的判别向量[19]。而 Friedman 等[20]考虑了对角阵估计。

$$\widetilde{\sum_w} = \operatorname{diag}(\widehat{\sigma_1^2}, \cdots, \widehat{\sigma_p^2}).$$

其中，$\widehat{\sigma_j^2}$ 是 $\widehat{\sum_w}$ 的第 j 个对角元，在本文中，我们使用的是 $\widetilde{\sum_w}$ 是对角线估计的情况，因为已经表明，当 $p > n$ 时，对 $\sum_w$ 使用对角线估计可以导致良好的分类结果[2]。

为解决正交约束，下面我们将用到这样一个引理。

引理 1：以下两个问题的解等价：

$$\max_{\beta_k} \quad \beta_k^T \widehat{\sum_b} \beta_k$$

$$\text{s. t.} \quad \beta_k^{\mathrm{T}} \widetilde{\sum_w} \beta_k \leqslant 1,$$

$$\beta_k^{\mathrm{T}} \widetilde{\sum_w} \beta_l = 0. \ (\forall\, l < k)$$

$$\max_{\beta_k} \quad \beta_k^{\mathrm{T}} \widehat{\sum_b^k} \beta_k$$

$$\beta_k^{\mathrm{T}}\widetilde{\sum}_w\beta_k<1.$$

式中，$\sum_b^k{}'=n^{-1}X^{\mathrm{T}}Y(Y^{\mathrm{T}}Y)^{-1/2}P_k^{\perp}(Y^{\mathrm{T}}Y)^{-1/2}Y^{\mathrm{T}}X$，$P_k^{\perp}$ 定义如下：$P_k^{\perp}=I$，对于 $k>1$，$P_k^{\perp}$ 是 $(Y^{\mathrm{T}}Y)^{-\frac{1}{2}}Y^{\mathrm{T}}X\hat{\beta}_i$，$i<k$ 张成的线性空间的补空间上的正交投影算子。

引理 1 的证明见参考文献[10]中的附录 A.2，当求出 $\widehat{\beta_1},\cdots,\widehat{\beta_k}$ 时，用 $\widehat{\sum}_b^{k+1}$ 替代原先的组间方差，重复之前的求解过程，直到得到所有的判别向量。

则原问题转变为：

$$\beta_k=\underset{\beta_k}{\operatorname{argmin}}\frac{\beta_k^{\mathrm{T}}\widetilde{\sum}_w\beta_k}{\beta_k^{\mathrm{T}}\widehat{\sum}_b\beta_k}.$$

求解上述问题相当于求解：

$$\beta_k=\underset{\beta_k}{\operatorname{argmin}}\beta_k^{\mathrm{T}}\widetilde{\sum}_w\beta_k-\lambda_1^{k-1}\beta_k^{\mathrm{T}}\widehat{\sum}_b\beta_k.$$

为得到稀疏的判别向量，对上式实施 l_0 惩罚

$$\beta_k=\underset{\beta_k}{\operatorname{argmin}}\beta_k^{\mathrm{T}}\widetilde{\sum}_w\beta_k-\lambda_1^{k-1}\beta_k^{\mathrm{T}}\widehat{\sum}_b\beta_k+\lambda_2\|\beta_k\|_0.$$

由于上述问题是非凸非光滑且 NP-难，我们转为用 l_q 范数逼近，上式转化为：

$$\beta_k=\underset{\beta_k}{\operatorname{argmin}}\beta_k^{\mathrm{T}}\widetilde{\sum}_w\beta_k-\lambda_1^{k-1}\beta_k^{\mathrm{T}}\widehat{\sum}_b\beta_k+\lambda_2\|\beta_k\|_q^q.(0<q\leqslant1)$$

其中，$\|\beta_k\|_q^q=\sum_j|\beta_j|^q$ 为了克服由 $\|\beta_k\|_q^q$ 引起的非光滑性，引入常规参数 ε，上式转为：

$$\beta_k=\underset{\beta_k}{\operatorname{argmin}}\beta_k^{\mathrm{T}}\widetilde{\sum}_w\beta_k-\lambda_1^{k-1}\beta_k^{\mathrm{T}}\widehat{\sum}_b\beta_k+\lambda_2\|\beta_k\|_{q,\in}^q.(0<q\leqslant1)$$

其中，$\|\beta_k\|_{q,\varepsilon}^q=\sum_{j=1}^p(\in+[\beta_k]_j^2)^{q/2}$，$[\beta_k]_j$ 表示 β_k 的第 j 个元素，当 ε 足够小的时候，$\|\beta_k\|_{q,\in}^q$ 是 $\|\beta_k\|_q^q$ 的一个较好的近似。

$$\min_{\beta}=\beta_k^{\mathrm{T}}\widetilde{\sum}_w\beta_k-\lambda_1^{k-1}\beta_k^{\mathrm{T}}\widehat{\sum}_b\beta_k+\lambda_2\|\beta_k\|_{q,\in}^q.(0<q\leqslant1)\tag{3.1}$$

3 算法及收敛性分析

3.1 算法

问题(3.1)的局部最小值点 $\beta_k^{\in,q}$ 需满足以下梯度方程

$$\left[\frac{\lambda_{2q}\beta_k^{\in,q}}{(\in+(\beta_k^{(t)})^2)^{1-\frac{y}{2}}}\right]_{1\leqslant k\leqslant p}+2\sum{}_w\beta_k^{\in,q}-2\widehat{\lambda_1}\widehat{\sum}_b\beta_k^{\in,q}=0.$$

令 $D_l=\operatorname{diag}\left[\frac{\lambda_{2q}}{(\in+(\beta_k^{(t)})^2)^{1-y/2}}\right]_{1\leqslant k\leqslant p}$，则上式可以写作：

$$(D_l+2\widetilde{\sum}_w-2\lambda_1\widehat{\sum}_b)\beta_k^{l+1}=0.\tag{4.1}$$

算法 1：

设置一个初始点 $\beta_1^0\in\mathbb{R}^p$，令 $l=0$

如果 $k>1$，定义一个正交投影矩阵 $P_k^{\perp}$，投影到正交于 $(Y^{\mathrm{T}}Y)^{-1/2}Y^{\mathrm{T}}X\hat{\beta}_i$ 的空间上，$\forall i<k$. 令：

$$\widehat{\sum}_w=n^{-1}\sum_{k=1}^k\sum_{i\in C_k}(x_i-\hat{\mu}_k)(x_i-\hat{\mu}_k)^{\mathrm{T}}.$$

$$\widetilde{\sum}_w=\operatorname{diag}(\hat{\sigma}_1^2,\cdots,\hat{\sigma}_p^2),P_1^{\perp}=I.$$

$$\widehat{\sum}_b^k=n^{-1}X^{\mathrm{T}}Y(Y^{\mathrm{T}}Y)^{-1/2}P_k^{\perp}(Y^{\mathrm{T}}Y)^{-1/2}Y^{\mathrm{T}}Y,\widehat{\sum}_b^1=\widehat{\sum}_b.$$

循环以下步骤：

(1)通过等式(4.1)利用 β_1^l 计算 β_1^{l+1}.

(2)$l=l+1$.

直到 $\|\beta^{(l)}-\beta^{(l+1)}\|^2$ 收敛时停止循环。

3.2 收敛性分析

为证明其收敛性，需要用到以下定理：

定理 1. 对于 $\in>0$，$\forall\alpha,\beta$ 并且 $0<q<1$，有

$$(\in+\alpha^2)^{\frac{q}{2}}-(\in+\beta^2)^{\frac{q}{2}}-\frac{q\beta(\alpha-\beta)}{(\in+\alpha^2)^{1-\frac{q}{2}}}\geqslant0.\tag{4.2}$$

定理 2. 令 $\in>0$，且 $\{\beta^{(l)}\}_{l=0}^{\infty}$ 是由算法 1 产生的序列，则序列 $\{L_q(\in,\beta^{(l)}\}_{l=0}^{\infty}$ 下降且收敛，即 $L_q(\in,\beta^{(l)})-L_q(\in,\beta^{(l+1)})\geqslant0$。

证明：

$$\begin{aligned}&L_q(\in,\beta^{(l)})-L_q(\in,\beta^{(l+1)})\\&=\lambda_2\sum_{j=1}^p(\in+(\beta_j^{(l)})^2)^{q/2}-\lambda_2\sum_{j=1}^p(\in+(\beta_j^{(l+1)})^2)^{q/2}+\\&\beta^{(l)\in}(\widetilde{\sum}_w-\lambda_1\widehat{\sum}_b)\beta^{(l)}-\\&\beta^{(l+1)\in}(\widetilde{\sum}_w-\lambda_1\widehat{\sum}_b)\beta^{(l+1)}\end{aligned}$$

$$\Rightarrow \lambda_2 \sum_{j=1}^{p} (\in + (\beta_j^{(l)})^2)^{q/2} - \lambda_2 \sum_{j=1}^{p} (\in + (\beta_j^{(l+1)})^2)^{q/2} + (\beta^{(l)} - \beta^{(l+1)})^{\mathrm{T}} (\overline{\sum}_w - \lambda_1 \widehat{\sum}_b)(\beta^{(l)} - \beta^{(l+1)}) + 2\beta^{(l)\in} (\widetilde{\sum}_w - \lambda_1 \widehat{\sum}_b)\beta^{(l+1)} - 2\beta^{(l+1)\in} (\widetilde{\sum}_w - \lambda_1 \widehat{\sum}_b)\beta^{(l)}$$

$$\Rightarrow \lambda_2 \sum_{j=1}^{p} (\in + (\beta_j^{(l)})^2)^{q/2} - \lambda_2 \sum_{j=1}^{p} (\in + (\beta_j^{(l+1)})^2)^{q/2} + (\beta^{(l)} - \beta^{(l+1)})^{\mathrm{T}} (\widetilde{\sum}_w - \lambda_1 \widehat{\sum}_b)(\beta^{(l)} - \beta^{(l+1)}) + 2(\beta^{(l)} - \beta^{(l+1)})^{\mathrm{T}} \left(\widetilde{\sum}_w - \lambda_1 \widetilde{\sum}_b\right)\beta^{(l+1)}.$$

由于上述梯度等式(4.1)可以写作：

$$\widetilde{\sum}_w - \lambda_1 \widehat{\sum}_b = D_l/2.$$

则：

$$2(\beta^{(l)} - \beta^{(l+1)})^{\mathrm{T}} (\widetilde{\sum}_w - \lambda_1 \widehat{\sum}_b)\beta^{(l+1)} = 2(\beta^{(l)} - \beta^{(l+1)})^{\mathrm{T}} (\mathrm{D}_l/2)\beta^{(l+1)} = -2(\beta^{(l)} - \beta^{(l+1)})^{\mathrm{T}} \left[\frac{\lambda_2 q}{2(\in + (\beta_j^{(l)})^2)^{1-q/2}}\right]_{1\leqslant j\leqslant p} \beta^{(l+1)} = -\sum_{j=1}^{p} \frac{\lambda_{2q}\beta_j^{(l+1)}(\beta_j^{(l)} - \beta_j^{(l+1)})}{(\in + (\beta_j^{(l)})^2)^{1-q/2}}$$

利用不等式(4.2)，可得：

$$L_q(\in, \beta^{(l)}) - L_q(\in, \beta^{(l+1)}) = \lambda_2 \sum_{j=1}^{p} \left[(\in + (\beta_j^{(l)})^2)^{q/2} - (\in + (\beta_j^{(l+1)})^2)^{q/2} - \frac{q\beta_j^{(l+1)}(\beta_j^{(l)} - \beta_j^{(l+1)})}{(\in + (\beta_j^{(l)})^2)^{1-q/2}}\right] + (\beta^{(l)} - \beta^{(l+1)})^{\mathrm{T}} (\widetilde{\sum}_w - \lambda_1 \widehat{\sum}_b)(\beta^{(l)} - \beta^{(l+1)}) \geqslant (\beta^{(l)} - \beta^{(l+1)})^{\mathrm{T}} (D_l/2)(\beta^{(l)} - \beta^{(l+1)}) \geqslant 0.$$

因此，序列$\{L_q(\in, \beta^{(l)})\}_{l=0}^{\infty}$是下降的，由于$L_q(\in, \beta^{(l)})$对于任意$\beta \in \mathbb{R}^p$都是非负的，因此序列同样是收敛的。

定理 3. 令$\in > 0$并且$0 < q < 1$则对任意初始点$\beta^0 \in \mathbb{R}^p$，由算法 1 产生的序列$\{\beta^{(l)}\}_{l=0}^{\infty}$都收敛到问题(3.1)的局部最小值点。

证明：

由定理 2，已知序列$\{L_q(\in, \beta^{(l)})\}_{l=0}^{\infty}$是下降的，且下界为 0，因此可以假设存在$M \geqslant 0$，使得$\lim_{l\to\infty} L_q(\in, \beta^{(l)}) = M$则有：

$$\| \beta_q^{(l)} \| \leqslant \| \beta^{(l)} \|_{q,\in} \leqslant L_q(\in, \beta^{(l)}) \leqslant M + 1.$$

对于足够大的l，有$\| \beta_q^{(l)} \|$是有界的，那么存在收敛子序列$\{\beta^{(l_i)}\}_{i=1}^{\infty}$和点$\hat{\beta} \in \mathbb{R}^p$使得$\lim_{l\to\infty} \beta^{(l_i)} = \hat{\beta}$。此外，由(4.1)可知，$\{\beta^{(l)}\}_{l=0}^{\infty}$的子序列$\{\beta^{(l_i+1)}\}_{i=1}^{\infty}$同样也是收敛的，即存在$\beta \in \mathbb{R}^p$，使得$\lim_{i\to\infty} \beta^{(l_i+1)} = \hat{\beta}$. 又由于$L_q(\in, \beta^{(l)})$在$\mathbb{R}^p$上是连续的，所以有$L_q(\in, \bar{\beta}) = L_q(\in, \hat{\beta})$，因此，由定理 2 可得

$$0 = L_q(\in, \bar{\beta}) - L_q(\in, \hat{\beta}) \geqslant (\bar{\beta} - \hat{\beta})^{\mathrm{T}} (D_l/2)(\bar{\beta} - \hat{\beta}) \geqslant 0.$$

即$(\bar{\beta} - \hat{\beta})^{\mathrm{T}} (D_l/2)(\bar{\beta} - \hat{\beta}) = 0$，则$\bar{\beta} = \hat{\beta}$，证毕。

4 总结与展望

4.1 总结

本章首先针对传统线性判别分析在高维情形下类内协方差矩阵奇异以及缺乏解释性这两个问题，建立了一种求解第k个判别成分的模型，该模型在原模型的基础上利用类内方差的对角估计矩阵代替原始类内协方差矩阵，克服了矩阵奇异的问题，同时将其投影到正交投影空间上，去掉了其正交约束，最后加入了l_q范数正则项，获得稀疏解，增强其解释性，并提出了求解该模型的迭代算法。最后给出了该算法的收敛性证明。

4.2 展望

本文在研究具有l_q正则项的稀疏判别分析的问题上，虽然详细讨论了求解第k个判别变量的模型和算法，但是并没有应用于实际问题中，对于是否可以有效改善其解释性，以及对线性判别分析的分类效果是否有积极作用缺乏说服力。因此可以将其应用于具体问题，观察其分类效果以及降维后数据的解释性是否得到有效提高。

利益冲突：作者声明无利益冲突。

参考文献(References)

[1] Taibi F, Akbarizadeh G, Farshidi E. Robust reservoir rock fracture recognition based on a new sparse feature

learning and data training method[J]. Multidimensional Systems and Signal Processing,2019,30(4):390-403.

[2] Sharifzadeh F,Akbarizadeh G,Kavian Y S. Ship Classification in SAR Images Using a New Hybrid CNN-MLP Classifier[J]. Journal of the Indian Society of Remote Sensing,2019,47(4):551-562.
https://doi.org/10.1007/s12524-018-0891-y

[3] Zhang Z Y,Wang J,Zha H Y. Adaptive manifold learning[J]. IEEE transactions on pattern analysis and machine intelligence,2012,34(2):253-265.
https://doi.org/10.1109/TPAMI.2011.115

[4] Hand D J. Classifier Technology and the Illusion of Progress[J]. Statistical Science,2006,21(1):1-15.

[5] Clemmensen L, Hastie T, Witten D, et al. Sparse discriminant analysis[J]. Techonometrics, 2011, 53(4): 406-413.
https://doi.org/10.1198/TECH.2011.08118

[6] Krzanowski W J,Jonathan P,McCarthy W V,et al. Discriminant Analysis with Singular Covariance Matrices: Methods and Applications to Spectroscopic Data[J]. Journal of the Royal Statistical Society. Series C(Applied Statistics),1995,44(1):101-115.

[7] Guo Y, Hastie T, Tibshirani R. Regularized linear discriminant analysis and its application in microarrays[J]. Biostatistics(Oxford,England),2007,8(1):86-100.

[8] Zou H,Hastie T,Tibshirani R. Sparse Principal Component Analysis[J]. Journal of Computational and Graphical Statistics,2006,15(2):265-286.

[9] Qiao Z H,Zhou L,Huang J Z. Sparse Linear Discriminant Analysis with Applications to High Dimensional Low Sample Size Data[J]. IAENG International Journal of Applied Mathematics,2009,39(1):48.

[10] Witten D M,Tibshirani R. Penalized classification using Fisher's linear discriminant[J]. Journal of the Royal Statistical Society. Series B, Statistical methodology, 2011,73(5):753-772.
https://doi.org/10.1111/j.1467-9868.2011.00783.x

[11] Shao J,Wang Y Z,Deng X W,et al. Sparse linear linear discriminant analysis by thersholding for high dimensional data[J]. The Annals of Statistics,2011,39(2): 1241-1265.
https://doi.org/10.2307/29783672

[12] Trevor H,Andreas B,Robert T. Penalized Discriminant Analysis[J]. The Annals of Statistics, 1995, 23(1): 73-102.

[13] Esedo $\bar{g}$lu S, Osher S J. Decomposition of images by the anisotropic Rudin-Osher-Fatemi model[J]. Communications on Pure and Applied Mathematics, 2004, 57(12):1609-1626.
https://doi.org/10.1002/cpa.20045

[14] Yin Q,Shu L. Sparse linear discriminant analysis via l_0 constraint[J]. Journal of University of Science and Technology of China,2022,52(08):21-27.

[15] Hoai A L T,Duy N P. DC programming and DCA for sparse optimal scoring problem[J]. Neurocomputing, 2016,186(1):170-181.
https://doi.org/10.1016/j.neucom.2015.12.068

[16] Li G Q,Duan X X,Wu Z Y,et al. Generalized elastic net optimal scoring problem for feature selection[J]. Neurocomputing,2021,447(447):183-195.
https://doi.org/10.1016/J.NEUCOM.2021.03.018

[17] Duintjer T J,Schlesinger P. Improving implementation of linear discriminant analysis for the high dimension small sample size problem[J]. Computational Statistics and Data Analysis,2007,52(1):423-437.
https://doi.org/10.1016/j.csda.2007.02.001

[18] Guo Y,Hastie T,Tibshirani R. Regularized linear discriminant analysis and its application in microarrays [J]. Biostatistics,2007,8(1):86-100.

[19] 尹祺. 基于 l_0 惩罚下的主成分分析与线性判别分析[D]. 合肥:中国科学技术大学,2022.
https://doi.org/10.27517/d.cnki.gzkju.2022.000185

[20] Friedman J H. Regularized Discriminant Analysis[J]. Journal of the American Statistical Association, 1989, 84(405):165-175.

[21] Mai Q,Zou H. A Note On the Connection and Equivalence of Three Sparse Linear Discriminant Analysis Methods[J]. Technometrics,2013,55(2):243-246.

Sparse Linear Discriminant Analysis Based on l_q Regularization

CHEN Jing, GAO Caixia*

(School of Mathematical Science, Inner Mongolia University, Hohhot 010021, China)

Abstract: Linear discriminant analysis plays an important role in feature extraction, data dimensionality reduction, and classification. With the progress of science and technology, the data that need to be processed are becoming increasingly large. However, in high-dimensional situations, linear discriminant analysis faces two problems: the lack of interpretability of the projected data since they all involve all p features, which are linear combinations of all features, as well as the singularity problem of the within-class covariance matrix. There are three different arguments for linear discriminant analysis: multivariate Gaussian model, Fisher discrimination problem, and optimal scoring problem. To solve these two problems, this article establishes a model for solving the kth discriminant component, which first transforms the original model of Fisher discriminant problem in linear discriminant analysis by using a diagonal estimated matrix for the within-class variance in place of the original within-class covariance matrix, which overcomes the singularity problem of the matrix and projects it to an orthogonal projection space to remove its orthogonal constraints, and subsequently an l_q norm regularization term is added to enhance its interpretability for the purpose of dimensionality reduction and classification. Finally, an iterative algorithm for solving the model and a convergence analysis are given, and it is proved that the sequence generated by the algorithm is descended and converges to a local minimum of the problem for any initial value.

Keywords: Linear discriminant analysis; sparse optimization; l_q norm; projection

DOI: 10.48014/fcpm.20230529001

Citation: CHEN Jing, GAO Caixia. Sparse linear discriminant analysis based on l_q regularization[J]. Frontiers of Chinese Pure Mathematics, 2023, 1(2): 31-38.

刊误更正

本刊2023年6月28日出版的第1卷1期21-30页，由兰州理工大学理学院郑亮撰写的论文“一类具有病毒感染的随机传染病模型的平稳分布”，漏标资助基金项目，现补充更正如下：“基金项目：甘肃省自然科学基金(资助号21JR7RA216)”。

特此更正！