文章编号:0253-2778(2020)02-0163-13

基于硬阈值惩罚函数的高维降秩回归

徐洪鸣

(中国科学技术大学管理学院统计与金融系,安徽合肥 230026)

摘要:为了解决多元回归问题中高维数据的复共线性,有一种方法是构造惩罚函数,来对估计矩阵的秩进行约束,它被称为降秩回归.为了得到更精确的估计,这里考虑用硬阈值函数做奇异值惩罚函数.通过局部线性近似方法,将原本的估计转换为可计算的模型.这个新的模型是可计算的且是连续的.在模拟和真实的数据集上与其他的模型进行实验比较分析,结果表明,这种估计在大部分情况下比一些常用的降秩估计拥有更高的精度.

关键词:硬阈值;奇异值分解;降秩回归模型;奇异值惩罚

中图分类号:O212.4 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2020.02.012

2010 Mathematics Subject Classification:Primary 62J07; Secondary 60F07

引用格式:徐洪鸣. 基于硬阈值惩罚函数的高维降秩回归[J]. 中国科学技术大学学报,2020,50(2):163-175.
XU Hongming. Reduced rank regression based on hard-thresholding singular value penalization[J]. Journal of University of Science and Technology of China, 2020,50(2):163-175.

Reduced rank regression based on hard-thresholding singular value penalization

XU Hongming

(*Department of Statistics and Finance, School of Management, University of Science and Technology, Hefei* 230026, *China*)

Abstract: Reduced rank estimation using penalty functions to restrict ranks of variety matrices is often used for solving the multi-collinearity of high-dimensional multivariate regression. Here a hard-thresholding singular value penalization was considered to get more efficient results. Through local linear approximate method, non-convex models were converted to computable ones. This model is computationally efficient, and the resulting solution path is continuous. Experiment results from simulation and public datasets show that this kind of reduced rank regression has better accuracy than some frequently-used ones in most situations.

Key words: hard-thresholding; singular value decomposition; reduced rank regression; singular value penalty

0 引言

目前,随着互联网和存储技术的大力发展,大数据的概念已普及,人们越来越倾向于使用海量数据来进行分析,这些数据常常有比较高的维度.如何从高维数据中准确而快速地得到有效信息成为当前统计学研究的热点问题.在处理高维回归问题时,由于数据维度的增加,模型变得更加复杂,因此往往有过拟合的风险,所以必须使用一些新方法来处理高维回归问题.

收稿日期:2019-01-12;**修回日期**:2019-06-02

作者简介:徐洪鸣,男,1995年生,硕士.研究方向:高维统计.E-mail:xuhongm@mail.ustc.edu.cn

降秩回归是一种处理高维多元线性回归问题的常见手段.这种方法的原理是通过降低估计矩阵的秩来降维从而易于估计.对于多响应问题来说,其中一种使用的方法是对 Frobenius 范数引出的损失函数添加一个惩罚函数来除去信息量较少的维度.由于因子的选择通常与估计的奇异值相关,且奇异值的个数就是矩阵的秩,因此这个惩罚函数通常是系数矩阵奇异值的函数.

2006 年,Yuan 等[1]提出了因子估计选择(factor estimation and selection,FES)方法,这种方法通过收缩和降维来进行估计,与其他的方法相比估计的性能较好.它的惩罚函数是系数矩阵奇异值的 L_1 惩罚,即 Ky Fan 1-范数.这个惩罚使得奇异值变得稀疏,同时实现了降秩和估计收缩[2].2010 年,Bunea 等[3]提出了秩选择准则(rank selection criterion,RSC)方法,它的惩罚函数是系数矩阵奇异值的 L_0 惩罚,这种方法的模型比前一个简单,因为它直接对秩进行惩罚,所以对变量个数的选择比较精准[4],对 n 较小的情况估计也较好.这两种方法分别对应软阈值函数和硬阈值函数.除此之外,还有一些比较复杂的模型及惩罚函数,如自适应核范数[4],平滑裁剪绝对偏差(smoothly clipped absolute deviation,SCAD)[5]和 elastic net[15]等.

Zheng 等[6]提出了一种新的惩罚函数,这种惩罚函数和 L_0 惩罚函数经过同样的变换都可以得到硬阈值函数,他们将这种惩罚函数命名为硬阈值惩罚函数并将它运用于单响应变量高维多元线性回归,在有关糖尿病的真实数据统计中获得较好的统计结果.

本文将这种惩罚函数用于多响应高维多元线性回归中,对系数矩阵的奇异值进行惩罚.它和 FES 及 RSC 都成功地规避了最小二乘法所带来的过拟合风险,并同时在一定程度上弥补了后两者的缺陷.对比 FES 方法,它对不同的因子附加不同的权重,使得估计的秩更加精准,而不像 FES 方法那样估计的秩通常偏大;对比 RSC,它给出的解的路径是连续的,并且在使用参数的选择时更加灵活,并且性能也会得到提升.

在本文中,首先,建立了一个标准的多响应多元线性回归模型,对它的 Frobenius 范数损失函数加上一个奇异值硬阈值惩罚.其次,定义了变量 $n^{-1/2}$ 标准化后的全局属性,发现这种惩罚函数与 RSC 方法的相似性(实际上它就是 RSC 方法的推广),并证明了预测变量与响应变量在选择时的损失是有界的,即满足 Oracle 不等式.接下来,由于这个惩罚函数非凸,利用局部线性近似估计方法来将模型变为可计算的形式.最后,将这个得到的模型进行模拟和真实数据实验,与之前提到的 FES 和 RSC 方法进行对比,发现该模型在大多数情况拥有更高的精度.

1 模型建立

对多元统计分析来说,通常定义是:它有 n 个样本,q 个响应变量 $y=(y_1,\cdots,y_q)'$ 和 p 个解释变量 $x=(x_1,\cdots,x_p)'$,然后模型为

$$Y=XB_0+E \tag{1}$$

式中,$Y=(y_1,\cdots,y_n)'$ 是一个 $n\times q$ 矩阵,$X=(x_1,\cdots,x_n)'$ 是一个 $n\times p$ 设计矩阵,B_0 为 $p\times q$ 系数矩阵,$E=(e_1,\cdots,e_n)'$ 是回归噪声,其中 e_i's 独立同分布于 $N(0,\Sigma)$,$r^*=\mathrm{rank}(B_0)$ 是因子矩阵 B_0 的秩,$r_x=\mathrm{rank}(X)$ 是设计矩阵 X 的秩.为了使所有的协变量都符合一个共同的尺度,我们假设每个协变量 x_j 都被调整为 L_2 范数为 $n^{1/2}$,与截距的常数协变量 1 相匹配.此时,有 $n^{-1}X^{\mathrm{T}}X=I_p$,这个过程被称为 $n^{-1/2}$ 标准化,因为 X 乘以 $n^{-1/2}$ 后是一个正交矩阵.

为了得到高维回归问题(1)比普通最小二乘估计的解 $B^{LS}=(X'X)^-X'Y$ 更好的解,可以通过估计取下面函数的极小值点来估计 B_0:

$$\frac{1}{2}J(B)+P(B) \tag{2}$$

式中,$B\in\mathbb{R}^{p\times q}$,$J(B)=\|Y-XB\|_F^2$ 是标准差,$\|\cdot\|_F$ 是 Frobenius 范数,$P(B)$ 是惩罚函数,其中至少包含一个调节惩罚函数的参数 λ.

降秩回归是对估计矩阵的秩做惩罚.由于在通常情况下认为估计矩阵的奇异值就是因子,同时矩阵秩的个数与奇异值个数相等,因此可以用奇异值惩罚来进行估计.对任意矩阵 $M\in\mathbb{R}^{m\times n}$ 都可以分解为 $M=UDV'$ 的形式,这被称为奇异值分解,其中 $U\in\mathbb{R}^{m\times m}$ 以及 $V\in\mathbb{R}^{n\times n}$ 都是正交矩阵,而且 $D\in\mathbb{R}^{m\times n}$ 是一个对角矩阵.其中,$D=\mathrm{diag}\{d_{11},d_{22},\cdots,d_{rr}\}$ 并且 $r=\mathrm{rank}(M)\leqslant\min(m,n)$,同时 $\{d_{11}^2,d_{22}^2,\cdots,d_{rr}^2\}$ 都是矩阵 $M'M$ 的特征值.奇异值惩罚函数是对估计矩阵的奇异值做惩罚,换句话说,这些惩罚函数 $P(B)$ 可以写成 $P(d_{11},d_{22},\cdots,d_{rr})$ 的形式.观察一些常见的惩罚函数,例如,L_0 惩罚函数的定义是

$$P_\lambda(B)=\lambda\cdot\mathrm{rank}(B)=\lambda\sum_{i=1}^{p\wedge q}I(\sigma_i(B)\neq 0),$$

其中 $I(\cdot)$ 是示性函数,而且 $\sigma_i(B)$ 表示的是矩阵

B 的第 i 大的奇异值；L_1 惩罚函数[1]

$$P_\lambda(B)=\lambda\ \|B\|_*=\lambda\sum_{i=1}^{p\wedge q}\sigma_i(B);$$

L_2 惩罚函数[7]

$$P_\lambda(B)=\lambda\ \|B\|_F^2=\lambda\sum_{i=1}^{p\wedge q}\sigma_i^2(B);$$

Schatten-b quasi-norm 惩罚函数[8]

$$P_\lambda(B)=\lambda\sum_{i=1}^{p\wedge q}\sigma_i^b(B);$$

它们都用的是奇异值惩罚函数.

所谓降秩回归，实际上就是控制估计 B 的秩 r. 秩 r 代表估计的因素的个数. 如果将 r 降得比较小，那么估计会避免过拟合，但太少的 r 会导致估计矩阵无法描述因素之间的关系，即欠拟合. 所以控制矩阵的秩在这里十分重要，同时由于离散性，控制的过程在某种意义上是非常不稳定的，某些数据的微小变化可能导致非常不同的估计.

考虑使如下最小二乘函数最小的 B：

$$Q(B)=\frac{1}{2n}\|Y-XB\|_F^2+\|p_\lambda(\sigma_i(B))\|_1 \tag{3}$$

式中，$\sigma_i(B)$是 B 中第 i 大的奇异值. 定义奇异值向量函数 SV 和奇异值向量 β：

$$\beta=SV(B)=(\sigma_1(B),\sigma_2(B),\cdots,\sigma_r(B)).$$

式中，r 是矩阵 B 的奇异值数目，然后定义

$$p_\lambda(\beta)=(p_\lambda(\sigma_1(B)),p_\lambda(\sigma_2(B)),\cdots,p_\lambda(\sigma_r(B)))'.$$

考虑惩罚函数 $p_\lambda(t)$，它的定义域是 $t\in[0,+\infty)$，值域是 $p_\lambda(t)\geqslant 0$，假设惩罚函数关于 t 和 λ 递增，而且 $p_\lambda(0)=0$. 这表明惩罚力度随着参数 λ 的大小和惩罚参数的增加而增加. 在 $n^{-1/2}$ 标准化后，设计矩阵 X 乘上 $n^{-1/2}$ 会变成正交矩阵，也就是说，$n^{-1}X'X=I_p$. 此时，$Q(B)$可以变为

$$Q(B)=\frac{1}{2}\|\hat{B}_{ols}-B\|_F^2+\|p_\lambda(\sigma_i(B))\|_1,$$

其中 $\hat{B}_{ols}=n^{-1}X'Y$ 是原回归的标准最小二乘估计.

本文将要讨论的奇异值惩罚函数 $p_\lambda(\sigma_i(B))$为

$$p_{H,\lambda}(t)=\frac{1}{2}\{\lambda^2-(\lambda-t)_+^2\},t\geqslant 0 \tag{4}$$

它与 L_0 惩罚 $p_{H_0,\lambda}(t)=2^{-1}\lambda^2 1_{\{t\neq 0\}}$，$t\geqslant 0$，十分接近，当 $t\geqslant\lambda$ 的时候，两种惩罚的力度是一样的. 事实上，通过计算 $\underset{X}{\operatorname{argmin}}\|X-B\|_F^2+\|p_\lambda(X)\|_1$ (X，B 均为 $n\times 1$ 向量)的解，可以发现它们的解都是$\{X_i=B_iI_{\{|B_i|>\lambda\}}\}$，即硬阈值函数，因此将这个新设定的惩罚函数命名为硬阈值惩罚函数. 这个结果说明本文给出的惩罚函数与 L_0 惩罚有着十分密切的关系.

2　相关定义、定理证明

2.1　模型成立条件

显而易见，硬阈值奇异值惩罚函数是一种连续的非凸惩罚函数，为了方便计算和研究，将它和与它接近的 L_0 惩罚函数联系起来. 先考虑 Y 是一维，即 $q=1$($y=X\beta+E$，$Q_0(B)=\|y-X\beta\|_F^2+p_\lambda(\beta)$，$p_\lambda$ 对 β 而不是 B 的奇异值作惩罚)的简单情况，可以得到如下引理.

引理 2.1　对硬阈值惩罚函数 $p_{H,\lambda}(t)$和 L_0 惩罚函数 $p_{H_0,\lambda}(t)$，沿着第 j 个坐标轴最小化 $Q_0(B)$，其中 $1\leqslant j\leqslant p$，对任意 p 维向量 β_j 第 j 个成分为零，考虑如下形式的坐标变量全局极小点 $\beta(z)=z1_{\{|z|>\lambda\}}$，有 $z=n^{-1}(y-X\beta_j)'X_j$.

证明　对硬阈值惩罚函数 $p_{H,\lambda}(t)$和 L_0 惩罚函数 $p_{H_0,\lambda}(t)$，它们关于 $\underset{X}{\operatorname{argmin}}\|X-B\|_F^2+\|p_\lambda(X)\|_1$ 的解是一样的$\{X_i=B_i1_{\{|B_i|>\lambda\}}\}$，考虑 $Q_0(B)$沿着第 j 个坐标轴最小化的表现形式为

$$\underset{\beta_j}{\operatorname{argmin}}\frac{1}{2n}\|y_j-X_j\beta_j\|_F^2+p_\lambda(\beta_j)=\underset{\beta_j}{\operatorname{argmin}}\|\beta_j-X'_jy_j\|_F^2,$$

所以这两种惩罚函数在 $q=1$ 时有相同的解 $\beta(z)=z1_{\{|z|>\lambda\}}$，$z=n^{-1}(y-X\beta_j)'X_j$.

接下来考虑 $Y\in\mathbb{R}^{n\times q}$的情况：将 B 分解为 $B=U\operatorname{diag}\{\sigma_1(B),\sigma_2(B),\cdots\}V'$，由于任何矩阵乘正交矩阵后奇异值不变，那么

$$\begin{aligned}Q(B)=&\frac{1}{2}\|\hat{B}_{ols}-U\operatorname{diag}\{\sigma_1(B),\sigma_2(B),\cdots\}V'\|_F^2+\\&\|p_{H,\lambda}(SV(B))\|_1=\\&\frac{1}{2}\|U_{ols}\operatorname{diag}\{\sigma_1(\hat{B}_{ols}),\cdots\}V_{ols}'-\\&U\operatorname{diag}\{\sigma_1(B),\sigma_2(B),\cdots\}V'\|_F^2+\\&\|p_{H,\lambda}(SV(B))\|_1=\\&\frac{1}{2}\|\operatorname{diag}\{\sigma_1(\hat{B}_{ols}),\cdots\}-\\&\operatorname{diag}\{\sigma_1(B),\sigma_2(B),\cdots\}\|_F^2+\\&\|p_{H,\lambda}(SV(B))\|_1.\end{aligned}$$

因此 $Q(B)$的最小值解 B 的奇异值向量 $SV(B)$同时是 $Q_0(\beta)=\|SV(\hat{B}_{ols})-\beta\|_2^2+\|p_{H,\lambda}(\beta)\|_1$ 的最小值解. 这个结果将 $Y\in\mathbb{R}^{n\times q}$与 $y\in\mathbb{R}^n$的情况联

系了起来.但两种估计是有差别的,显然不能保证两个估计量相同.

众所周知,高共线性通常与大规模数据集相关.高的共线性会导致估计的稳定性较差,甚至在回归问题中模型可辨识性的损失.更具体地说,可能存在另一个与 β_0 不同的 p 维向量 β_1,使得 $XU\beta_1V'$ 与 $XU\beta_0V'$ 几乎相同,当维数 p 与样本大小 n 相比较大时,为了保证模型的可辨识性和减小模型不稳定性,必须控制稀疏模型的大小,因为从几何角度可以清楚地看出协变量之间的共线性随着维数的增加而增加.这种思想在文献[9]中被用于稀疏恢复问题,即 (1) 的无噪声情况.为了保证 β_0 的可辨识性,引入了火种的概念,计为 spark(X),用于设计矩阵 X,其定义为最小数量的 τ,使得存在 X 的一个线性相关的 τ 列子集,特别地,只要 $s<\mathrm{spark}(X)/2$,β_0 是唯一的,它提供了模型可辨识性的基本条件.

由于变量要在存在噪声的情况下选择,扩展鲁棒火种概念如下:

定义 2.1 对一个 $n\times p$ 并且有界 c 的设计矩阵 X,有鲁棒火种 $M=\mathrm{rspark}_c(X)$. 其中 c 定义为考虑最小的 τ 使得 $n^{-1/2}X$ 存在 τ 列构成的子矩阵的奇异值具有小于给定的正常数 c.

定义 2.1 中 $M=\mathrm{rspark}_c(X)$ 的一种等价表示是最大的符合下列不等式的 τ:

$$\min_{\|\delta\|_0<\tau,\|\delta\|_2=1} n^{-1/2}\|X\delta\|_2\geqslant c \tag{5}$$

这个不等式提供了稀疏模型的共线性的自然约束,它由文献[9]提出来解决稀疏恢复问题,将用于定理 2.1 的证明.由 s 的定义,有 $s<M/2$,并在下面的条件 2.2 中,它的约束条件与文献[10]的条件相似.约束特征值条件假设(1)和 L_0 范数约束 $\|\delta\|_0<\tau$ 取代 L_1 范数约束 $\|\delta J_0^c\|_1\leqslant c_0\|\delta J_0\|_1$(对于某些正常数 c_0,其中 $J_0\subset\{1,\cdots,p\}$ 以及 $\|J_0\|\leqslant s'$),其中 J_0^c 是 J_0 的补集,而且 δ_A 表示 δ 在给定集合 A 中由具有指数的部分组成的子向量.鲁棒性条件 $s<M/2$ 需要 (1) 中至少有 $\tau\geqslant 2s+1$.因为 L_0 范数约束通常定义在 L_1 范数约束时 $s'=2s$ 的子集上,因此鲁棒性条件通常弱于特征值约束条件.容易得出鲁棒火种 $\mathrm{rspark}_c(X)$ 与 c 负相关,而且当 $c\to 0+$ 时趋于 spark(X).因此 M 一般可以是任何不大于 $n+1$ 正整数.考虑正则估计量 $\hat{B}$ 关于坐标子空间的并集 $\mathbb{S}_{M/2}=\{B\in\mathbb{R}^{p\times q}:\|\beta\|_0<M/2\}$,因此,本文定义的全局最优解为

$$\hat{B}=\underset{B\in\mathbb{S}_{M/2}}{\operatorname{argmin}}\,Q(B) \tag{6}$$

式中,$Q(B)$ 的定义见式(3).

为了便于技术分析,类比文献[6]定义约束条件,给出以下两个正则条件.

条件 2.1 对所有 $E_{ij}s$,它们服从均值为 0 的正态分布,方差最大为 σ^2.

条件 2.2 因为 $s<M/2$,$s=o(n)$,以及

$$b_0=\min_{j\in\mathrm{supp}(\beta_0)}|\beta_{0,j}|>$$

$$(\sqrt{16/c^2}\wedge 1)c^{-1}\left(\frac{\sigma(\sqrt{p}+\sqrt{q})}{\sqrt{n}}+\frac{c_2\sqrt{\ln n}}{\sqrt{n}}\right)\sqrt{2s+1},$$

其中 M 是 X 的鲁棒火种,由定义 2.1,它的界为 c,$c_2\geqslant\sqrt{2}\sigma$ 是一个正常数.

条件 2.1 是线性回归模型的标准结构.用高斯误差分布的假设来简化技术参数.实际上这个条件并不是必要的,定理 2.1 对于其他的误差分布也会成立,例如文献[11]的 2.1 节中的例子.再考虑条件 2.2,前半部分的 $s<M/2$ 对包含鲁棒火种的大小为 s 的真实模型设置稀疏约束,即需要这样一个鲁棒火种条件来确保模型的可辨识性.通常假设样本大小 n 与真协变量 s 的比发散,即,$s=o(n)$,因此可以获得 β_0 的一致估计.

2.2 模型整体性质和收敛效果

基于引理 2.1,正则化参数 λ 对硬阈值惩罚函数 $p_{H,\lambda}(t)$ 和 L_0 惩罚 $p_{H_0,\lambda}(t)$ 的约束力度相同,因此选择适当的足够大的正则化参数 λ 来抑制所有的噪声协变量并保留重要的协方差参数可以确保模型选择一致性.这种方法被证明是在对模型的一致性和阈值回归 Oracle 不等式定理有效.

这个有效性的证明见文献[12].

定理 2.1 假设条件 2.1 和 2.2 成立并且

$$c^{-1}\left(\frac{\sigma(\sqrt{p}+\sqrt{q})}{\sqrt{n}}+\frac{c_2\sqrt{\ln n}}{\sqrt{n}}\right)\sqrt{2s+1}<$$

$$\lambda<b_0(1\wedge\sqrt{c^2/2}).$$

然后对硬阈值惩罚函数 $p_{H,\lambda}(t)$ 和 L_0 惩罚函数 $p_{H_0,\lambda}(t)$,考虑对一些正常数 $c_2'>\sqrt{2}\sigma$,这两种惩罚函数的正则估计 $\hat{B}$ 在 $\hat{B}\geqslant 1-\mathrm{e}^{-\frac{n\theta^2(\sqrt{p}+\sqrt{q})^2}{2\sigma^2}}-\mathrm{e}^{-\frac{2n\theta_0^2r^*}{\sigma^2}}$ 的概率下同时满足下面条件:

(a)模型变量选择一致性:

$$\mathrm{rank}(\hat{B})=\mathrm{rank}(B_0);$$

(b)预测损失:

$$n^{-1/2}\|X(\hat{B}-B_0)\|_F\leqslant\frac{4\sqrt{2}s\sigma}{c\sqrt{n}}+\frac{2c_2'\sqrt{s\ln n}}{c\sqrt{n}};$$

(c)估计损失:

$$\| \hat{B} - B_0 \|_m \leqslant \frac{4\sqrt{2}s^{(1/2+1/m)}\sigma}{c^2\sqrt{n}} + \frac{2c'_2 s^{1/m}\sqrt{\ln n}}{c^2\sqrt{n}};$$

对任意 $m \in [1,2]$ 成立且当 $m=\infty$ 与 $m=2$ 时，这个估计损失上界是相同的.

定理 2.1 的证明见附录.

将式(1)两边同时乘以 $(X'X)^{-1}X'$，得到

$$(X'X)^{-1}X'Y = B_0 + (X'X)^{-1}X'E,$$

因此，为了保证估计的可靠性，估计 B 的奇异值要大于误差 $(X'X)^{-}X'E=\frac{1}{n}X'E$ 的奇异值. 现在讨论 $\frac{1}{n}X'E$. $\frac{1}{n}X'E$ 是一个 $p\times q$ 矩阵而且 $d_1(\frac{1}{n}X'E)$ 是 $\frac{1}{n}X'E$ 的最大奇异值. 可以通过下面的方法来估计这个值.

定义 2.2　(亚高斯随机变量)定义一个随机变量 X 具有亚高斯性当存在亚高斯矩 $K>0$ 使得

$$P(|X|>t) \leqslant 2\mathrm{e}^{-t^2/K^2},\ t>0 \tag{7}$$

亚高斯随机变量的例子包括正态随机变量、± 1 值和一般所有有界随机变量.

设 A 为 $n_1\times n_2$ 随机矩阵，它的每个元素为亚高斯随机变量，且各项独立，均值为 0，亚高斯矩有限，那么它具有如下性质：

$$P(d_1(A) > C(\sqrt{n_1}+\sqrt{n_2})+t) \leqslant 2\mathrm{e}^{-ct^2} \tag{8}$$

式中，$d_1(A)$ 是 A 的最大奇异值，C，c 均为正常数. 因为 X 是一个 $n^{-1/2}$ 标准化设计矩阵，而 $\frac{1}{n}X'E$ 是一个 $p\times q$ 随机矩阵，它是各项独立，均值为 0，亚高斯矩有限的亚高斯随机变量可参考文献[12]，因为 $\frac{1}{n}X'E_{ij}\sim N(0,\sigma^2/n)$，那么有

$$P(d_1(\frac{1}{n}X'E) > C(\sqrt{p}+\sqrt{q})+t) \leqslant \mathrm{e}^{-ct^2/2} \tag{9}$$

式中，$C=\sigma/\sqrt{n}$ 而且 $c=n/\sigma^2$，也就是说

$$P(d_1(\frac{1}{n}X'E) > \frac{\sigma}{\sqrt{n}}(\sqrt{p}+\sqrt{q})+t) \leqslant \mathrm{e}^{-nt^2/(2\sigma^2)} \tag{10}$$

当定理 2.1(a)成立时，此时估计的系数矩阵 B 的秩为 r^*，B 通过奇异值分解为 $B=UDV'$，式(1)等价于

$$YV = XUD + EV,$$

由于系数矩阵 B 的秩为 r^*，通常认为 $r^*<n$(否则样本量小于因子量，根本无法估计)，可以得到

$$Y_0 = X_0\hat{B}_0 + E_0,$$

其中，$Y_0\in\mathbb{R}^{n\times r}$ 是 YV 的子矩阵，$X_0\in\mathbb{R}^{n\times r}$ 是 XU 的子矩阵，$\hat{B}_0\in\mathbb{R}^{r\times r}$ 是 D 的子矩阵，$E_0\in\mathbb{R}^{n\times r}$ 是 EV 的子矩阵. 这样就像是把 D 中为 0 的奇异值项对应的行或列摘去，得到一个浓缩过的式子. 定义两种情况

$$\epsilon = \{ d_1(\frac{1}{n}X'E) \leqslant (\frac{\sigma}{\sqrt{n}}+\theta)(\sqrt{p}+\sqrt{q})\} \tag{11}$$

和

$$\epsilon' = \{ d_1(\frac{1}{n}X'_0E_0) \leqslant (\frac{\sigma}{\sqrt{n}}+\theta_0)(\sqrt{2r^*})\} \tag{12}$$

其中 θ 和 θ_0 都是正常数，有

$$P(\epsilon^c) \leqslant \mathrm{e}^{-\frac{n\theta^2(\sqrt{p}+\sqrt{q})^2}{2\sigma^2}} \tag{13}$$

$$P(\epsilon'^c) \leqslant \mathrm{e}^{-\frac{n\theta_0^2*2r^*}{\sigma^2}} \tag{14}$$

这两个关于 $P(\epsilon^c)$ 和 $P(\epsilon'^c)$ 的不等式十分相似，令式(9)中的 $t=\theta(\sqrt{p}+\sqrt{q})$ 得到式(13)，在式(14)中 X'_0E 是一个 $r^*\times r^*$ 矩阵，其中 $r^*=\mathrm{rank}(\hat{B})$，令式(10)中的 $t=2\theta_0(\sqrt{r^*})$ 即可得式(14). 定理 2.1 是基于情况 $\epsilon_1=\epsilon\bigcap\epsilon'$(概率至少为 $1-\mathrm{e}^{-\frac{n\theta^2(\sqrt{p}+\sqrt{q})^2}{2\sigma^2}}-\mathrm{e}^{-\frac{2n\theta_0^2r^*}{\sigma^2}}$，这个概率当 $n\to\infty$ 时趋于 1).

为了确保这个概率当 $n\to\infty$ 时趋于 1，必须使 $P(\epsilon^c)$ 和 $P(\epsilon'^c)$ 趋于 0. 所以可以得到 θ 和 θ_0 的约束条件：$\theta\geqslant c_2\frac{\sqrt{\ln n}}{\sqrt{n}(\sqrt{p}+\sqrt{q})}$ 和 $\theta_0\geqslant c'_2\frac{\sqrt{\ln n}}{\sqrt{n}(\sqrt{2r^*})}$. 同时，注意到 $r^*\leqslant s'=2s$，所以 $\sqrt{r^*}\leqslant\sqrt{2s}$.

利用上述正则化参数 λ 的选择，正则化估计关于 Oracle 估计量的预测损失在对数因子 $(\ln n/n)^{1/2}$ 以内，这被称为真正的基础稀疏模型的最小二乘估计. 定理 2.1 还建立了正规估计量在 L_q 估计损失当 $q\in[1,2]\cup\infty$ 的 Oracle 不等式. 这些结果同时以很高的概率收敛到一个与样本大小 n 有关的多项式，其中维度因子 p，q 允许随样本大小 n 快速增长但被约束在与 λ 有关的范围.

定理 2.1 是基于引理 2.1 的，它们都有一个共同的特征，即它们的结论对硬阈值惩罚函数和 L_0 惩罚函数同样适用. 这是因为条件 2.1 和 2.2 从硬阈值惩罚函数 $p_{H,\lambda}(t)$ 和 L_0 惩罚函数 $p_{H_0,\lambda}(t)$ 得到的估计是大致相同的. 事实上，在不同的预测和变量选择损失下，它们的 Oracle 不等式的收敛速度是

渐近等价的.

3 硬阈值惩罚函数回归求解方法的实现

3.1 局部线性近似方法

因为硬阈值回归是非凸惩罚回归,因此本文采用局部线性近似(local linear approximate,LLA)方法来简化这一问题.文献[13]提出了局部线性近似方法,该方法相对于局部二次逼近和扰动局部二次逼近具有三个显著的优点.首先,在局部线性近似方法中,我们不必删除任何小的系数或选择扰动的大小,以避免数值不稳定.其次,局部线性近似方法是最佳的凸最小化-最大化(MM)算法,从而通过 MM 算法的提升性质证明了局部线性近似方法算法的收敛性[14].第三,局部线性近似方法通过连续惩罚产生稀疏估计,由此我们可以得到一步局部线性近似方法.计算上,一步局部线性近似方法减轻了迭代算法中的计算负担,并克服了潜在的局部极大值问题中最大化非凹惩罚似然.

在惩罚函数的基础上,利用局部线性逼近得到新的惩罚函数的统一算法,它的原理是

$$p_\lambda(|\beta_j|)\approx p_\lambda(|\beta_j^{(0)}|)+p'_\lambda(|\beta_j^{(0)}|)(|\beta_j|-|\beta_j^{(0)}|),$$

其中,$\beta_j\approx\beta_j^{(0)}$.然后我们考虑一步局部线性近似方法[13],因为在稀疏的线性回归模型中,局部线性近似方法容易得到一步估计量(即第二步估计与第一步相同).为了简单起见,把初始估计 $\beta^{(0)}$ 定义为普通最小二乘估计.然后通过一步估计得到

$$\beta^{(1)}=\operatorname{argmin}\frac{1}{2}\|y-X\beta\|^2+n\sum_{j=1}^{p}p'_\lambda(|\beta_j^{(0)}|)|\beta_j| \tag{15}$$

此外,我们将该思想推广到新模型,并得到新的一步估计量:

$$B^{(1)}=\operatorname{argmin}\frac{1}{2}\|Y-XB\|^2+n\sum_{j=1}^{p}p'_\lambda(|\beta_j^{(0)}|)|\beta_j| \tag{16}$$

式中,$\beta=(\sigma_1(B),\sigma_2(B),\cdots,\sigma_r(B))$,而且 $p'_\lambda(t)=(\lambda-t)1_{\{\lambda>t\}}$,$t\geqslant 0$.同样,我们可以进一步定义 k 步局部线性近似,但通常情况下因为我们的估计几乎都是低秩的,它意味着 k 步局部线性近似与一步局部线性近似结果基本相同.现在,我们用局部线性近似方法将问题转化为一个加权 lasso 惩罚回归模型.

3.2 加权 lasso 惩罚回归

在前面的部分中,我们发现,如果能够解决加权 lasso 惩罚回归,那么可以解决原本的阈值回归问题.文献[4]提出了一种基于加权 L_1 核范数,在高维多元回归中同时降维和系数估计的新方法.它结合了两个主要优点.首先,这种方法建立了 L_0 和 L_1 奇异值惩罚方法之间的桥梁,它可以被视为类似于文献[13,16-17]开发的自适应 lasso 单变量回归问题.其次,惩罚 XB,而不是 B,这使得降秩估计问题可以显式和高效地解决,这种思路也被文献[18-19]在跟踪回归问题中使用.

首先,我们将加权 lasso 看成是其奇异值的加权和:

$$\|B\|_{*\omega}=\sum_{i=1}^{p\wedge q}\omega_i\sigma_i(B) \tag{17}$$

当 $B\in\mathbb{R}^{p\times q}$ 以及 $\sigma_i(B)$ 表示的是 B 的第 i 个奇异值.然后由于毕达哥拉斯定理,考虑最小化下面最小二乘惩罚的加权 lasso 惩罚回归:

$$\frac{1}{2}\|Y-XB\|_F+\|XB\|_{*\lambda\omega} \tag{18}$$

当然权重 $\{\lambda\omega_i\}$ 必须是非负的和非降的.这表明,如果惩罚函数 $\sum_{i=1}^{p\wedge q}p_\lambda(\beta_i)=\lambda\sum_{i=1}^{p\wedge q}\omega_i\beta_i$ 以及 $0\leqslant\omega_1\leqslant\cdots\leqslant\omega_{p\wedge q}$,加权 lasso 惩罚函数回归有一个固定的最小值.而且当 $\omega_1=\omega_{p\wedge q}$,新的回归将成为文献[1]中的一个凸惩罚函数回归.文献[4]中定义 $\Phi_{\lambda\omega}(B)=\operatorname{diag}\{\sigma_i(B)-\lambda\omega\}_+$,而且找到一个极小值 $\widehat{B}_s^{\lambda\omega}$:

$$\widehat{B}_s^{\lambda\omega}=B_{\mathrm{LS}}VD^-\Phi_{\lambda\omega}(D)V' \tag{19}$$

式中,$B_{\mathrm{LS}}=(X'X)^-X'Y$ 是 B 的最小二乘估计,而且 UDV' 是 XB_{LS} 的奇异值分解.

现在观察新的加权 lasso 惩罚函数回归,因为 $\sigma_1(B)\geqslant\cdots\geqslant\sigma_r(B)$,$\omega_i=(\lambda-\sigma_i(B))I_{\{\lambda>\sigma_i(B)\}}$,所以它符合 $0\leqslant\omega_1\leqslant\cdots\leqslant\omega_r$,因此,我们可以用文献[4]中提到的方法来解这个加权 lasso 惩罚回归.

使用前两节提到的方法,我们得到算法如下:

算法 3.1 硬阈值奇异值惩罚函数法的算法

输入:数据矩阵 $X\in\mathbb{R}^{n\times p}$,$Y\in\mathbb{R}^{n\times q}$.

输出:控制参数 λ,估计矩阵 $\widehat{B}$.

①计算 $B_{\mathrm{LS}}=(X'X)^-X'Y$,将 B_{LS} 奇异值分解为 $B_{\mathrm{LS}}=UDV'$,其中 $D=\operatorname{diag}\{d_1,d_2,\cdots\}$.

②计算 $M=XB_{\mathrm{LS}}$,将其奇异值分解为 $M=\widehat{U}\widehat{D}\widehat{V}'$,而且 $\widehat{D}=\operatorname{diag}\{D_1,D_2,\cdots\}$.

③令 $\Psi^{\lambda}(\hat{D})=\mathrm{diag}\{D_j-(\lambda-d_j)_+\}_{n\times q}$. 计算 $\hat{B}(\lambda)=B_{\mathrm{LS}}\hat{V}\hat{D}^{-}\Phi^{\lambda}(\hat{D})\hat{V}'$.

④求出使 $\mathrm{tr}\{(Y-X\hat{B}(\lambda))'(Y-X\hat{B}(\lambda))\}$最小的 λ,输出这个 λ 和对应的 $\hat{B}(\lambda)$.

4　数据实验

4.1　模拟实验

我们通过进行模拟试验来比较本文和常用的降秩估计的预测精度和估计精度,在 MATLAB 环境中实现.

秩选择方法[3]:

$$\frac{1}{2}\|Y-XB\|_F+\lambda^2 r(B) \tag{20}$$

其估计记为 $B_H^{(\lambda)}$.

因子惩罚函数方法[1]:

$$\frac{1}{2}\|Y-XB\|_F+\lambda\|B\|_* \tag{21}$$

其估计记为 $B_N^{(\lambda)}$.

本文提出的硬阈值奇异值惩罚函数导出的估计记为 $B_T^{(\lambda)}$.

考虑文献[3]中提到的模型. 系数矩阵 B_0 构造方法为 $B_0=bB_1B_2'$,其中 $b>0$, $B_1\in\mathbb{R}^{p\times r^*}$, $B_2\in\mathbb{R}^{q\times r^*}$ 而且所有的 B_1 和 B_2 都是从 $N(0,1)$中随机取样. 考虑了模型尺寸的两种情况,分别是 $p,q<n$ 或 $p,q>n$,对这两种情况分别定义为模型 1 和模型 2. 在模型 1 中,设定 $n=100$, $p=q=25$, $r^*=10$. 矩阵 X 通过将其 N 行生成为服从 $N(0,\Gamma)$的随机样本来构造,其中 $\Gamma=(\Gamma_{ij})_{p\times q}$ 而且 $\Gamma_{ij}=\rho^{|i-j|}$, $0<\rho<1$. 在模型 2 中,设定 $n=40$, $p=q=50$ 以及 $r^*=10$, $r^x=20$. 矩阵 X 的构造方法是 $X=X_0\Gamma^{1/2}$,其中 Γ 按之前的方法构造, $X_0=X_1X_2$, $X_1\in\mathbb{R}^{n\times r^x}$, $X_2\in\mathbb{R}^{r^x\times p}$, 而且所有 X_1,X_2 都随机取样自 $N(0,1)$.

数据矩阵 Y 的构造方法是 $Y=XB_0+E$,其中 E 中的向量都是 $N(0,1)$中的随机样本. 每个模拟模型都有样本大小 n,预测维度 p,响应维度 q,模型的真实秩 r^*,设计矩阵的秩 r^x,相关系数 $\rho\in\{0.1,0.5,0.9\}$和信号强度 $b\in\{0.05,0.1,0.3\}$. 对每种参数设置重复实验 500 次.

对每种估计方法,模型精度的比例由所有 500 次运行均方误差的平均值来测量,其中,估计精度 $\mathrm{Est}(B)=100\|B_0-\hat{B}\|_F^2/(pq)$, 预测精度 $\mathrm{Pred}(B)=100\|XB_0-X\hat{B}\|_F^2/(nq)$. 表 1 和 2 总结了模拟结果,并列出了模型 1 和 2 的各种降秩估计方法在模拟数据集的精度,为了对比显示降秩估计的精度,这里还给出标准最小二乘的估计 $\hat{B}_{\mathrm{LS}}=(X'X)^{-}X'Y$ 的精度,显然相比几种降秩回归估计,用标准最小二乘得到的估计的精度很低.

表 1　模型 1 下几种估计方法的预测和估计均方误差比较

Tab. 1　Comparison of the estimation and prediction errors between kinds of estimators using Model 1

ρ	b		$\hat{B}_H^{(\lambda)}$	$\hat{B}_N^{(\lambda)}$	$\hat{B}_T^{(\lambda)}$	$\hat{B}_{\mathrm{LS}}^{(\lambda)}$
0.9	0.05	Est	2.518 8	1.896 6	1.934 9	12.471 7
		Pred	12.108 4	10.394 4	9.824 1	25.104 7
	0.1	Est	5.432 6	3.890 8	3.988 8	12.381 6
		Pred	15.980 1	14.366 2	13.214 0	24.953 1
	0.3	Est	6.851 6	6.601 5	6.437 0	12.383 2
		Pred	16.982 5	19.344 0	16.532 2	24.963 3
0.5	0.05	Est	1.170 7	0.835 7	0.832 8	2.232 6
		Pred	16.561 6	12.130 9	12.789 7	25.048 3
	0.1	Est	1.344 1	1.187 6	1.130 8	2.217 6
		Pred	17.431 2	16.734 9	15.620 8	25.081 9
	0.3	Est	1.248 8	1.487 7	1.289 4	2.213 0
		Pred	16.170 6	19.731 6	16.759 2	24.970 4

续表 1

ρ	b		$\widehat{B}_H^{(\lambda)}$	$\widehat{B}_N^{(\lambda)}$	$\widehat{B}_T^{(\lambda)}$	$\widehat{B}_{LS}^{(\lambda)}$
0.1	0.05	Est	0.866 3	0.650 9	0.644 1	1.372 1
		Pred	17.242 2	13.501 8	13.260 0	24.908 5
	0.1	Est	0.892 0	0.849 9	0.803 8	1.386 7
		Pred	17.271 0	16.975 2	15.913 1	25.083 4
	0.3	Est	0.828 7	1.002 9	0.855 4	1.377 9
		Pred	16.149 9	19.550 2	16.676 2	24.983 7

表 2 模型 2 下几种估计方法的预测和估计均方误差比较

Tab. 2 Comparison of the estimation and prediction errors between kinds of estimators using Model 2

ρ	b		$\widehat{B}_H^{(\lambda)}$	$\widehat{B}_N^{(\lambda)}$	$\widehat{B}_T^{(\lambda)}$	$\widehat{B}_{LS}^{(\lambda)}$
0.9	0.05	Est	1.061 9	0.979 6	0.981 7	2.393 5
		Pred	26.942 4	25.336 7	23.957 0	49.712 8
	0.1	Est	3.501 5	3.329 6	3.350 8	4.572 4
		Pred	30.672 4	31.728 6	29.244 6	50.137 7
	0.3	Est	27.372 4	27.744 6	27.544 3	28.513 4
		Pred	31.635 9	41.843 3	32.402 9	49.838 7
0.5	0.05	Est	0.892 7	0.858 9	0.863 1	1.160 4
		Pred	30.428 6	28.648 1	27.572 4	49.857 7
	0.1	Est	2.130 0	2.108 6	2.178 4	3.424 2
		Pred	31.424 8	34.285 5	31.194 8	49.824 2
	0.3	Est	27.472 2	27.038 7	26.832 8	27.792 2
		Pred	32.265 5	42.614 2	32.542 9	49.953 7
0.1	0.05	Est	0.854 4	0.855 2	0.845 9	1.071 9
		Pred	31.113 4	29.018 4	27.876 1	50.067 5
	0.1	Est	3.117 6	3.214 0	3.094 5	3.348 4
		Pred	30.942 1	34.786 6	31.413 1	49.649 4
	0.3	Est	27.212 7	27.419 5	27.299 0	27.510 0
		Pred	32.029 0	41.995 0	32.361 7	50.257 6

考虑到对每种方法的公平性，调节参数 λ 使得所有的估计矩阵 $B_N^{(\lambda)}$，$B_H^{(\lambda)}$ 和 $B_T^{(\lambda)}$ 恰好取到估计的最小均方误差，也就是说调节参数 λ 来取得最接近的估计.

通过从表 1 和 2 中模型 1 和 2 得到的结果，可以得出以下结论：在模型 1 中硬阈值回归惩罚函数估计 $B_T^{(\lambda)}$ 的估计精度与预测精度优于因子惩罚函数估计 $B_N^{(\lambda)}$ 的和秩选择估计 $B_H^{(\lambda)}$ 的；在模型 2 中硬阈值回归惩罚函数估计 $B_T^{(\lambda)}$ 的与秩选择估计 $B_H^{(\lambda)}$ 的估计精度和预测精度要优于因子惩罚函数估计 $B_N^{(\lambda)}$ 的，但后者当 b 较小时精准度比较高. $B_T^{(\lambda)}$ 在模型 1 中表现最好，在模型 2 中与 $B_H^{(\lambda)}$ 不相上下，总的来说 $B_T^{(\lambda)}$ 是一种精准度较高的估计方法. 而 $B_{LS}^{(\lambda)}$ 是没有加惩罚函数的结果，明显它的误差要大于三种加了惩罚函数的估计的. 同时我们将 $\rho = 0.1, 0.5, 0.9$ 时不同的 b 和估计方法得到的 Est 和 Pred 求均值（$B_{LS}^{(\lambda)}$ 误差太大不好比较），这样就可以得到图 1～4，这四幅图也可以印证我们之前得到的结论. 在表 1 和 2 中，我们还发现当 b（信噪比）较小并且 ρ（相关系数）较大时，$B_N^{(\lambda)}$ 的表现比 $B_T^{(\lambda)}$ 稍好，但此时 $B_N^{(\lambda)}$ 的计算

量很大而且常常会过拟合[3].

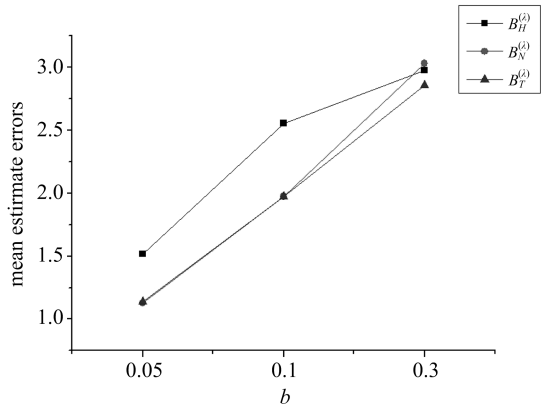

图 1　模型 1 下不同估计方法的平均估计误差关于 b 的变化曲线

Fig. 1　Curves of mean estimate errors from different estimators and b in Model 1

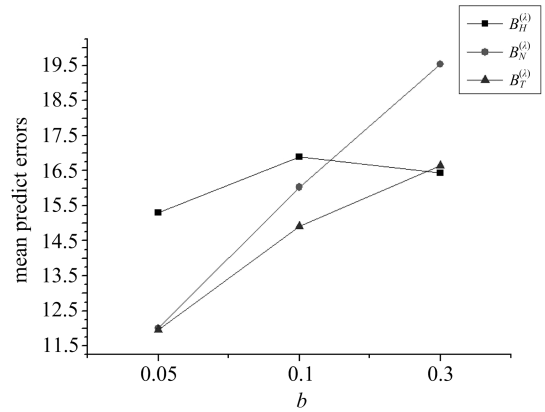

图 2　模型 1 下不同估计方法的平均预测误差关于 b 的变化曲线

Fig. 2　Curves of mean predict errors from different estimators and b in Model 1

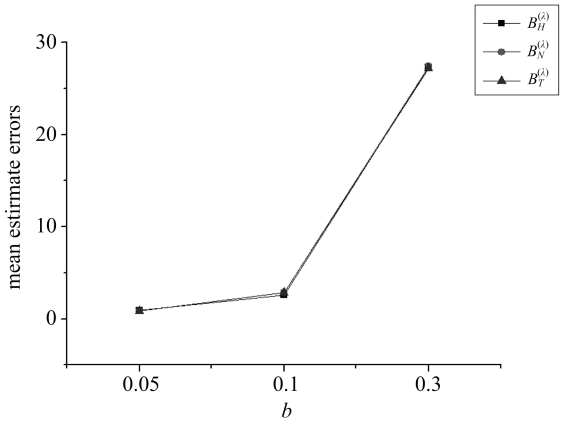

图 3　模型 2 下不同估计方法的平均估计误差关于 b 的变化曲线

Fig. 3　Curves of mean estimate errors from different estimators and b in Model 2

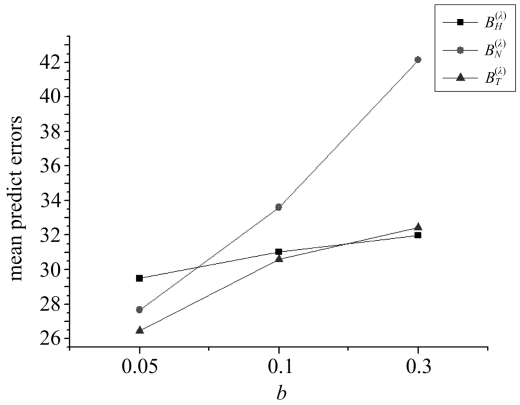

图 4　模型 2 下不同估计方法的平均预测误差关于 b 的变化曲线

Fig. 4　Curves of mean predict errors from different estimators and b in Model 2

4.2　实际应用

用 breast cancer 数据集[20]来进行真实数据测试，由该数据集由 $n=89$ 组基因表达数据和比较基因组杂交数据组成. 它也可以在 R 包 PMA 中找到. 它的详细描述参见文献[21]. 先前的研究已经表明，某些类型的癌症的特征是 RNA 异常的改变. 生物学上，在 RNA 变异上回归基因表达谱是有意义的，因为与给定基因相对应的 DNA 部分的扩增或缺失可能导致该基因表达的相应增加或减少. 同样，由 RNA 来预测 DNA 也是有意义的，因为所得到的预测模型可以识别功能相关的 RNA 变化. 尝试这两种方法，即设定 1：指定染色体的 RNA 变异作为预测因子和同一染色体的基因表达谱作为响应；设定 2：逆转预测因子和反应的角色. 发现，在设定 1 中，这三种方法没有一种能适当地拟合数据，秩选择准则甚至可能无法得到结果. 降秩模型在设定 2 下给出了更好的结果. 因此，只考虑设定 2 的结果. 现在设定预测变量和响应变量. 按照刚才得到的结论，指定基因表达谱 DNA 作为同一染色体的预测变量和 RNA 变异作为响应变量，因此 $p=19672$ 以及 $q=2194$. 当然，本文不使用那样庞大的数据总量，而使用来自其中一个染色体(21 号)的数据.

对 21 号染色体进行分析，共有 44 个 DNA 数据，227 个 RNA 数据，即，$p=227$ 以及 $q=44$. 通过下面的实验来比较各种降秩方法. 通过十折交叉法来实验，同时这些数据被随机分成两组测试数据，其中 $n_{\text{train}}=79$(几乎所有数据用作训练) 以及 $n_{\text{test}}=10$，而且正好 $n_{\text{train}}>\min(p,q)$(符合模型 1) 以及 $n_{\text{test}}<\min(p,q)$(符合模型 2). 首先对训练数据进

行估计,通过观察其均方预测误差

$$\mathrm{MSPE}=\|Y_{\mathrm{test}}-X_{\mathrm{test}}\widehat{B}\|_F^2/(q^{n_{\mathrm{test}}})$$

选出使均方误差最小的 λ. 然后使用测试数据来评估每个估计方法的预测性能. 这个随机的分裂过程重复 100 次,所得平均均方预测误差如表 3 所示.

表 3 21 号染色体数据训练数据与预测数据的误差比较

Tab. 3 Comparison of the errors between three kinds of reduced rank estimators in train data and test data about Chromosome 21

	MSPE		
	$\widehat{B}_H^{(\lambda)}$	$\widehat{B}_N^{(\lambda)}$	$\widehat{B}_T^{(\lambda)}$
train data	0.688 3	0.700 1	0.679 6
test data	0.932 4	0.850 2	0.818 7

从表 3 中可以看出,本文提出的新方法的估计 $\widehat{B}_T^{(\lambda)}$ 在训练集和测试集的误差小于 $\widehat{B}_H^{(\lambda)}$ 和 $\widehat{B}_N^{(\lambda)}$ 的,说明它的估计精度较优. 同时,可以发现对三种方法都有测试集误差高于训练集误差,这说明对降秩回归来说 n 较大的情形估计较好,即对符合模型 1 的数据的预测优于符合模型 2 的数据的.

5 结论

本文提出了一种使用硬阈值惩罚函数的降秩高维多元回归模型. 该模型使用了一种新的惩罚函数,通过付出提升计算量的代价来得到更高的估计精度,此外,还采用了局部线性近似方法来得到这个非凸问题的解. 模拟数据集与 breast cancer 真实数据集的实验结果表明,硬阈值惩罚函数方法在精度上优于一些常见的降秩方法,在需要更高精度的实际应用中可能会有更好的表现. 实际上,还有很多的惩罚函数现在还没有应用到降秩回归中,它们也许比硬阈值惩罚函数拥有更高的精度,今后我们将进一步对它们在降秩回归中的应用展开研究.

参考文献(References)

[1] YUAN M, EKICI A, LU Z, et al. Dimension reduction and coefficient estimation in multivariate linear regression [J]. Journal of the Royal Statistical Society, 2010, 57 (3): 329-346.

[2] NEGAHBAN S N, WAINWRIGHT M J. Estimation of (near) low-rank matrices with noise and high-dimensional scaling[J]. International Conference on Machine Learning, 2010, 39(2): 823-830.

[3] BUNEA F, SHE Y, WEGKAMP M H. Optimal selection of reduced rank estimators of high-dimensional matrices [J]. Annals of Statistics, 2010, 39 (2): 1282-1309.

[4] CHEN K, DONG H, CHAN K S. Reduced rank regression via adaptive nuclear norm penalization [J]. Biometrika, 2013, 100(4): 901-920.

[5] FAN J, LI R. Variable selection vianonconvave penalized likelihood and its oracle properties[J]. Publications of the American Statistical Association, 2001, 96 (456): 1348-1360.

[6] ZHENG Z, FAN Y, LV J. High dimensional thresholded regression and shrinkage effect[J]. Journal of the Royal Statistical Society B, 2014, 76(3) : 627-649.

[7] HOERL A E, KENNARDR W. Ridge regression: Biased estimation for nonorthogonal problems [J]. Technometrics, 2000, 42(1): 80-86.

[8] ROHDE A, TSYBAKOV A B. Estimation of high-dimensional low-rank matrices[J]. Annals of Statistics, 2011, 39(2): 887-930.

[9] DONOHO D L, ELAD M. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization[J]. Proceedings of the National Academy of Sciences of the United States of America, 2003, 100 (5): 2197-2202.

[10] BICKEL P J, RITOV Y, TSYBAKOV A B. Simultaneous analysis of lasso and Dantzig selector[J]. Annals of Statistics, 2008, 37(4): 1705-1732.

[11] FAN J, LV J. Nonconcave penalized likelihood with NP-dimensionality[J]. IEEE Transactions on Information Theory, 2011, 57(8): 5467-5484.

[12] REINSEL G C, VELU R P. Multivariate Reduced-Rank Regression[M]. New York: Springer, 1998: 369-370.

[13] ZOU H, LI R. One-step sparse estimates innonconcave penalized likelihood models [J]. Annals of Statistics, 2008, 36(4): 1509-1533.

[14] LANGE K, HUNTER D R, YANG I. Optimization transfer using surrogate objective functions[J]. Journal of Computational and Graphical Statistics, 2000, 9 (1): 1-20.

[15] ZOU H, HASTIE T. Regularization and variable selection via the elastic net[J]. J Roy Statist Soc Ser B, 2005, 67 (2): 301-320.

[16] TIBSHIRANI R. Regression shrinkage and selection via the lasso [J]. Journal of the Royal Statistical Society, 2011, 73(3): 273-282.

[17] HUANG J, HOROWITZ J L, MA S. Asymptotic properties of bridge estimators in sparse high-dimensional

regression models[J]. Annals of Statistics, 2008, 36(2): 587-613.

[18] KLOPP O. Rank penalized estimators for high-dimensional matrices[J]. Electronic Journal of Statistics, 2011, 5(2011): 1161-1183.

[19] KOLTCHINSKII V, LOUNICI K, TSYBAKOV A B. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion[J]. Annals of Statistics, 2011, 39(5): 2302-2329.

[20] WITTEN D M, TIBSHIRANI R, HASTIE T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis[J]. Biostatistics, 2009, 10(3): 515-534.

[21] CHIN K, DEVRIES S, FRIDLYAND J, et al. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies [J]. Cancer Cell, 2006, 10(6): 529-541.

[22] VON NEUMANN J. Some matrix-inequalities and metrization of matrix-space[J]. Tomsk Univ Rev, 1937, 11(1): 286-300.

附录

定理 2.1 的证明

A.1　证明需要的引理

引理 A.1　(Von Neumann 迹不等式)考虑两个矩阵:$A, B \in \mathbb{R}^{n_1 \times n_2}$. 它们的奇异值组成的向量为 $\sigma(A)=\{a_1, a_2, \cdots\}$, $\sigma(B)=\{b_1, b_2, \cdots\}$,那么有

$$\mathrm{tr}(A'B)=\langle A, B\rangle \leqslant \langle \sigma(A), \sigma(B)\rangle = a_1 * b_1 + a_2 * b_2 \cdots$$

引理 A.1 的证明见文献[22].

A.2　模型变量选择一致性

第一步,我们假设 $\hat{\beta}=(\sigma_1(\hat{B}), \cdots, \sigma_r(\hat{B}))$, $\beta_0=(\sigma_1(B_0), \cdots, \sigma_r(B_0))$ $(r=\min(p, q))$,任何非零分量的真实回归系数向量 β_0 或全局最优解 $\hat{\beta}$ 比 λ 大,这说明 $\| p_\lambda(\hat{\beta}) \|_1 = \lambda^2 \| \hat{\beta} \|_0 / 2$, $\| p_\lambda(\beta_0) \|_1 = s\lambda^2/2$. 所以, $\| p_\lambda(\hat{\beta}) \|_1 - \| p_\lambda(\beta_0) \|_1 = (\| \hat{\beta} \|_0 - s)\lambda^2/2$. 所以,我们用 δ 代替 $\hat{\beta}-\beta_0$. 直接计算

$$Q(\hat{B}) - Q(B_0) = \frac{1}{2n}(\mathrm{tr}\{(Y - X\hat{B})'(Y - X\hat{B})\} - \mathrm{tr}\{(Y - XB_0)'(Y - XB_0)\}) + \| p_\lambda \sigma_i(\hat{B}) \|_1 - \| p_\lambda \sigma_i(B_0) \|_1.$$

其中,

$$\mathrm{tr}\{(Y - X\hat{B})'(Y - X\hat{B})\} - \mathrm{tr}\{(Y - XB_0)'(Y - XB_0)\} =$$
$$\mathrm{tr}\{Y'Y - \hat{B}'X'(X\hat{B} + E) - (X\hat{B} + E)'X\hat{B} + \hat{B}'X'X\hat{B}\} -$$
$$\mathrm{tr}\{Y'Y - B_0'X'(XB_0 + E) - (XB_0 + E)'XB_0 + B_0'X'XB_0\} =$$
$$\mathrm{tr}\{(\hat{B} - B_0)'X'X(\hat{B} - B_0)\} - 2\mathrm{tr}\{E'X(\hat{B} - B_0)\}.$$

考虑到 $B'=\{\beta_1', \cdots, \beta_n'\}$,然后 $\mathrm{tr}\{B'X'XB\} = \sum_{i=1}^{n} \| X\beta_i \|_2$. 考虑到鲁棒火种 $M=\mathrm{rspark}_c(X)$是使如下不等式成立的最大的 τ:

$$\min_{\| \delta \|_0 < \tau, \| \delta \|_2 = 1} n^{-1/2} \| X\delta \|_2 \geqslant c.$$

所以有

$$n^{-1/2}\mathrm{tr}\{B'X'XB\} \geqslant c \ (\max(\| \beta_i \|_0) < M, \| B \|_F = 1).$$

而且因为引理 A.1,有

$$\mathrm{tr}(A'B) \leqslant \sum \sigma_i(A)\sigma_i(B).$$

由于一个矩阵第一个奇异值最大,所以我们得到条件更松的另一个迹不等式 $\mathrm{tr}(A'B) \leqslant d_1(A)\sum \sigma_i(B)$,那么有

$$n^{-1}\mid \mathrm{tr}\{E'X(\hat{B}-B_0)\}\mid = n^{-1}\mid \mathrm{tr}\{E'X\hat{B}\}-\mathrm{tr}\{E'XB_0\}\mid \leqslant$$

$$n^{-1}d_1(X'E)\mid \sum(\sigma_i(\hat{B})-\sigma_i(B_0))\mid \leqslant (\frac{\sigma}{\sqrt{n}}+\theta)(\sqrt{p}+\sqrt{q})\parallel\delta\parallel_1 \leqslant$$

$$(\frac{\sigma}{\sqrt{n}}+\theta)(\sqrt{p}+\sqrt{q})\parallel\delta\parallel_0^{1/2}\parallel\delta\parallel_2.$$

将这些式子合并可以得到

$$Q(\hat{B})-Q(B_0)\geqslant 2^{-1}c^2\parallel\delta\parallel_F^2-(\frac{\sigma}{\sqrt{n}}+\theta)(\sqrt{p}+\sqrt{q})\parallel\delta\parallel_0^{1/2}\parallel\delta\parallel_2+(\parallel\hat{\beta}\parallel_0-s)\lambda^2/2.$$

所以，

$$2^{-1}c^2\parallel\delta\parallel_F^2-(\frac{\sigma}{\sqrt{n}}+\theta)(\sqrt{p}+\sqrt{q})\parallel\delta\parallel_0^{1/2}\parallel\delta\parallel_2+(\parallel\hat{\beta}\parallel_0-s)\lambda^2/2\leqslant 0.$$

现在定义 t 等于$(\frac{\sigma}{\sqrt{n}}+\theta)(\sqrt{p}+\sqrt{q})$，重新整理这些公式，得到

$$\{c\parallel\delta\parallel_2-\frac{t}{c}\parallel\delta\parallel_0^{1/2}\}^2-\frac{t^2}{c^2}\parallel\delta\parallel_0+(\parallel\hat{\beta}\parallel_0-s)\lambda^2\leqslant 0.$$

可以得出

$$(\parallel\hat{\beta}\parallel_0-s)\lambda^2\leqslant\frac{t^2}{c^2}\parallel\delta\parallel_0.$$

定义 $k=\parallel\hat{\beta}\parallel_0=\mathrm{rank}(\hat{B})$，令 $\parallel\delta\parallel_0=\parallel\hat{\beta}-\beta_0\parallel_0\leqslant k+s$. 因此

$$(k-s)\lambda^2\leqslant\frac{t^2}{c^2}\parallel k+s\parallel_0.$$

整理 k 和 s 的关系，我们得到 $k\{\lambda^2-\frac{t^2}{c^2}\}\leqslant s\{\lambda^2+\frac{t^2}{c^2}\}$. 所以

$$k\leqslant s(\lambda^2+\frac{t^2}{c^2})/(\lambda^2-\frac{t^2}{c^2})=s\{1+\frac{2t^2}{\lambda^2c^2-t^2}\}<s+1.$$

因此，$\parallel\hat{\beta}\parallel_0\leqslant s$.

第二步的做法是基于第一步，假设 $\parallel\beta_0\parallel_0<\parallel\hat{\beta}\parallel_0$；那么丢失的真实相关系数的个数 $k=\mid\parallel\beta_0\parallel_0-\parallel\hat{\beta}\parallel_0\mid\geqslant 1$. 所以我们有 $\parallel\hat{\beta}\parallel_0\geqslant s-k$ 和 $\parallel\delta\parallel_0\leqslant\parallel\hat{\beta}\parallel_0+\parallel\beta_0\parallel_0\leqslant 2s$. 综合以上结论，有

$$Q(\hat{B})-Q(B_0)\geqslant 2^{-1}c^2\parallel\delta\parallel_2^2-\sqrt{2s}t\parallel\delta\parallel_2-k\lambda^2/2.$$

对所有 $j\in\mathrm{supp}(\beta_0)\backslash\mathrm{supp}(\hat{\beta})$，有 $\mid\delta_j\mid=\mid\beta_{0,j}\mid\geqslant b_0$. 所以，$\parallel\delta\parallel_2\geqslant b_0\sqrt{k}$，综上，有

$$4^{-1}c^2\parallel\delta\parallel_2\geqslant 4^{-1}c^2b_0\sqrt{k}\geqslant 4^{-1}c^2b_0>\sqrt{2s}t.$$

因此

$$Q(\hat{B})-Q(B_0)\geqslant 4^{-1}c^2\parallel\delta\parallel_2^2-k\lambda^2/2\geqslant 4^{-1}c^2kb_0^2-k\lambda^2/2>0.$$

因为 $\lambda<b_0c/\sqrt{2}$. 所以，有 $\mathrm{supp}(\beta_0)\subset\mathrm{supp}(\hat{\beta})$，即 $s\leqslant\parallel\hat{\beta}\parallel_0$. 结合 $\parallel\hat{\beta}\parallel_0\leqslant s$，我们得到 $\parallel\hat{\beta}\parallel_0=s$，即 $\mathrm{rank}(\hat{\beta})=\mathrm{rank}(\beta_0)$.

A.3 预测和估计损失

$X(\hat{B}-B_0)$的 Frobenius 范数

$$\parallel X(\hat{B}-B_0)\parallel_F=\sqrt{\sum\sigma^2(X(\hat{B}-B_0))}=\sqrt{\mathrm{tr}\{(\hat{B}-B_0)'X'X(\hat{B}-B_0)\}}.$$

我们考虑情况 $\epsilon_1=\epsilon\bigcap\epsilon'$（$\epsilon$ 和 ϵ' 见式(11)和(12)），有 $\parallel\delta\parallel_0\leqslant s$，又由于上面刚证明的 A.2，且由于 Cauthy-Schwarz 不等式，有

$$\mid n^{-1}E'_0X_0\delta\mid\leqslant d_1(n^{-1}E'_0X_0)\mid\sum(\sigma_i(\hat{B})-\sigma_i(B_0))\mid\leqslant$$

$$(\frac{\sigma}{\sqrt{n}}+\theta_0)(\sqrt{2r^*})\ \|\delta\|_1\leqslant\sqrt{s}(\frac{\sigma}{\sqrt{n}}+\theta_0)(\sqrt{2r^*})\|\delta\|_2.$$

根据 A.2 中给出的 $\|\hat{\beta}\|_0=s$,所以

$$Q(\hat{B})-Q(B_0)=2^{-1}\|n^{-1}X(\hat{B}-B_0)\|_F^2-\mathrm{tr}\{n^{-1}E'X(\hat{B}-B_0)\}+1/2(\|\hat{\beta}\|_0-s)\lambda^2\geqslant$$

$$2^{-1}c^2\|\delta\|_2^2-d_1(n^{-1}E'X)\|\delta\|_2\geqslant 2^{-1}c^2\|\delta\|_2-(\frac{\sigma}{\sqrt{n}}+\theta_0)(\sqrt{s})(\sqrt{2r^*})\|\delta\|_2.$$

从全局最优性 β 有 $2^{-1}c^2\|\delta\|_2-(\frac{\sigma}{\sqrt{n}}+\theta_0)(\sqrt{s})(\sqrt{2r^*})\leqslant 0$,其中 L_2 估计和 L_∞ 估计的边界

$$\|\hat{\beta}-\beta_0\|_\infty\leqslant\|\hat{\beta}-\beta_0\|_2=\|\delta\|_2\leqslant\frac{2}{c^2}(\frac{\sigma}{\sqrt{n}}+\theta_0)(\sqrt{s})(\sqrt{2r^*})\leqslant\frac{4\sqrt{2}s\sigma}{c^2\sqrt{n}}+\frac{2c'^2_2\sqrt{s\ln n}}{c^2\sqrt{n}}.$$

对 L_m 估计损失当 $1\leqslant m\leqslant 2$,应用 Holder 不等式得到

$$\|\hat{\beta}-\beta_0\|_m=(\sum_{j=1}^{n}|\delta_j|^m)^{1/m}\leqslant(\sum_{j=1}^{n}|\delta_j|^2)^{1/2}(\sum_{\delta_j\neq 0}1^{2/(2-m)})^{1/m-1/2})=$$

$$\|\delta\|_2\|\delta\|_0{}^{1/m-1/2}\leqslant 2s^{1/m}\frac{1}{c^2}(\frac{\sigma}{\sqrt{n}}+\theta_0)(\sqrt{2r^*})\leqslant\frac{4s^{(1/2+1/m)}\sigma}{c^2\sqrt{n}}+\frac{2c'^2_2s^{1/m}\sqrt{\ln n}}{c^2\sqrt{n}}.$$

最后,证明了 Oracle 预测损失的界.因为 $\hat{B}$ 是全局最优解,结合 A.2 的证明对 ϵ_1,有

$$2^{-1/2}n^{-1/2}\mathrm{tr}\{(\hat{B}-B_0)'X'X(\hat{B}-B_0)\}\leqslant\{n^{-1}\mathrm{tr}\{E'X(\hat{B}-B_0)\}-(\|\hat{\beta}\|_0-s)\lambda^2/2\}^{1/2}\leqslant$$

$$d_1(n^{-1}X'_0E_0)\ \|\delta\|_1\leqslant\sqrt{2s}\ \frac{1}{c}(\frac{\sigma}{\sqrt{n}}+\theta_0)(\sqrt{2r^*})\leqslant\frac{2\sqrt{2}s\sigma}{c\sqrt{n}}+\frac{c'^2_2\sqrt{2s\ln n}}{c\sqrt{n}}.$$

这样就完成了 $n^{-1/2}$ 标准化设计矩阵情况下的证明.