密级：公开　　中图分类号：0121.1

☑全日制　☐非全日制

浙江工商大學

硕士学位论文

（专业学位）

论文题目： **基于先验稀疏框架的 Lasso 回归模型的研究及其运用**

作者姓名： 曹金磊

专业学位类别： 应用统计硕士

专业学位领域： 应用统计

研究方向： 数理统计

指导教师： 王伟刚

提交日期：2020 年 1 月

Dissertation Submitted to Zhejiang Gongshang University

for Master's Degree of Economics

Research and Application of Lasso Regression Model Based on Prior Sparse Frame

Author: **Cao Jinlei**

Major: **Statistics**

Supervisor: **Prof. Wang Weigang**

Jan. 2020

School of Statistics and Mathematics

Zhejiang Gongshang University

Hangzhou, 310018, P. R. China

摘　要

随着近些年大数据时代的崛起，数据的分析和处理在社会科学、信息科学、遗传学、生物学、医学和金融学等各大科学领域都日益受到重视，因此如何从海量数据中对其本质特征进行提取成为一个重要的研究方向。并且就海量数据而言，我们如何建立合适的数据模型并挖掘出特征数目较少但信息相对全面的数据来进行分析就成为每个数据科学家需要面对的问题。

在众多数据模型中，Lasso模型是一种可以有效处理高维数据，且不损失相应精准度的模型。它是一种典型的变量选择方法，即可以通过设置阈值，限制参数总和的大小来压缩部分较小变量为0并剔除多余的变量。对于传统的回归模型，Lasso回归模型及其改进的模型能够很好的解决其在变量选择方面的问题，因此Lasso方法及其改进的方法在统计学研究中受到很大的重视。

本文针对Lasso回归模型，提出了一种新的Lasso改进方法，即将先验信息融入Lasso回归模型。本文将其称为基于先验稀疏框架的Lasso回归模型。首先，本文介绍了Lasso回归模型在回归问题上相较于其他模型的优势，并深入浅出的阐述了Lasso回归模型的求解算法及其具有的良好性质。其次，本文引入稀疏框架的概念与Lasso 回归模型进行结合，介绍了一种更为一般的Lasso回归框架。其可以转化为已知的多种Lasso变形。再次，先验信息作为回归分析特征自由属性的一部分，其本身不能用于描述研究对象，所以需要通过特有的方式将特征本身的先验信息应用到模型中去。因此本文引用一般的Lasso回归框架，给出了基于先验稀疏框架的Lasso回归模型的定义，并且从理论上给出了相应的求解算法和性质。最后，本文通过对多组模拟数据和实证数据进行分析，分析结果表明在具有先验信息的情况下，基于先验稀疏框架的Lasso 回归模型相较于普通Lasso 回归模型具有更好的性能。

关键词: Lasso模型；先验信息；稀疏框架；坐标下降法

Abstract

With the rise of the era of big data in recent years, the analysis and processing of data has received increasing attention in various scientific fields such as social sciences, information science, genetics, biology, medicine and finance, so how to deal with massive data The extraction of its essential features has become an important research direction. And in terms of massive data, how to build a suitable data model and mine data with a small number of features but relatively comprehensive information for analysis is a problem that every data scientist needs to face.

Among the many data models, the Lasso model is a model that can effectively process high latitude data without losing the corresponding accuracy. It is a typical variable selection method, which can compress some small variables to 0 and eliminate redundant variables by setting the threshold and limiting the size of the parameter sum. For the traditional regression model, the Lasso regression model and its improved model can solve the problem of variable selection well. Therefore, the Lasso method and its improved method have received great attention in statistical research.

In this paper, a new Lasso improvement method is proposed for the Lasso regression model. The prior information is incorporated into the Lasso regression model. This paper refers to the Lasso regression model based on the prior sparse framework. First of all, this paper introduces the advantages of the Lasso regression model on the regression problem compared with other models, and explains the algorithm of the Lasso regression model and its good properties. Secondly, this paper introduces the concept of sparse framework and Lasso regression model, and introduces a more general Lasso regression framework. It can be converted into a variety of known Lasso variants. Again, the prior information is part of the free attribute of the regression analysis feature, which itself cannot be used to describe the research object, so it is necessary to apply the prior information of the first feature itself to the model in a unique way. Therefore, this paper cites the general Lasso regression framework, and gives the definition of Lasso regression model based on prior sparse framework, and theoretically gives the corresponding algorithm and properties. Finally, this paper analyzes multiple sets of simulation data and empirical data. The results show that the Lasso regression model based on prior sparse framework

has better performance than the ordinary Lasso regression model with prior information.

Keywords: Lasso model; Prior information; Sparse frame; Coordinate descent method

目 录

第1章　绪论

1.1　选题的背景及研究意义

1.1.1　选题背景

随着科技的进步，收集数据的技术也有了迅猛的发展，因此如何有高效、快速地从数据中挖掘出有价值的信息越来越受到人们的关注。在大数据时代，通常采用统计学和机器学习的方法来对数据进行处理。它在很多领域中均具有非常显著的影响，最常见的有人工智能、机器识别、遗传学、医学、金融和市场等。虽然应用层面具有不同的领域，但是大数据问题却又具有几个共同的特点：(1) 数据量非常大，包含几百万甚至达到亿级别的训练量；(2) 数据的维度很高，通常每个样本均具有详细信息来记录其特征；(3) 大规模的数据通常以分布式的方式储存或收集。在机器学习的算法中，对最新算法的要求既能够解决数据的复杂性又能够采用平行或分布式的方法处理大数据。考虑通常的线性回归情形，我们有数据集X 的大小为$n \times p$，y作为响应变量，假定$y = X\beta + \varepsilon$, 其中噪声是高斯分布，符合$\varepsilon \sim N(0, \sigma^2 I)$。通常的最小二乘估计是通过最小化残差平方和而的得到的。下面是求解线性回归的目标函数

$$min \frac{1}{2} \|y - X\beta\|_2^2. \tag{1-1}$$

虽然这个最小二乘估计有许多好的性质，但仍然不能满足许多情况下数据分析的要求，主要是存在两个问题。首先是预测精度的问题。最小二乘估计虽然是无偏估计，但是它的方差在自变量线性相关程度高时通常较大，可以通过将某些稀疏压缩到0来改进这个预测精度。其次，是模型的可解释性，对于自变量个数很多的情况下，我们总是希望确定一个较小的变量模型来表现出最好的效果。有两种方法可以对最小二乘估计进行改进，分别是子集选择和岭回归，但是他们都有各自的缺点。子集选择虽然使模型可解释，但却使模型变得不稳定，这是它的离散型程序决定的——回归系数要么是被保留要么就是简单地从模型中剔除，这很可能使得观测数据的一个微小变动就导致要选择一个新的模型，从而影响了预测的准确性。而岭回归是一个连续型的方法，它缩小了回归系数，而且没有简单的抛掉那个变量，模型比较稳定，但正是由于它没有让任何一个回归系数减少到0，使得模型中变量太多，模型的解释性不好。Lasso 是1996年Tibshirani提出的一种方法，Lasso 的目标是为了解决下面这个问题

$$min\frac{1}{2}\|y-X\beta\|_2^2+\lambda\|\beta\|_1. \tag{1-2}$$

Lasso包含平方误差的求和以及L1正则化项。这种方法用模型系数的绝对值函数作为惩罚来压缩模型系数，使一些回归系数变小，甚至还使一些绝对值较小的系数压成了0.Lasso 方法集合了子集选择法和岭回归各自的优点，既可以将一些因子的系数压缩到0，又保持了连续性，并为模型提供了更好的解释性和预测性。因此目前经常利用Lasso来处理共线性问题和变量选择问题。

本文稀疏框架在此指的是稀疏框架矩阵，其是用于描述稀疏编码来表示真实的β，其目标函数为

$$min\frac{1}{2}\|y-X\beta\|_2^2+\lambda\|T\beta\|_1. \tag{1-3}$$

其中T是指定的稀疏矩阵，当T为单位矩阵时，该模型既为Lasso模型。对于T矩阵进行转化可以转化成基本所有的改进后的Lasso模型。

1.1.2 理论意义

在统计建模中，Lasso方法能够很好的进行变量选择和变量的压缩，目前无论是在学术研究中还是实际应用中都成为了行之有效的方法。因此通过稀疏框架和Lasso 方法很自然的进行结合，利用已知的先验信息对Lasso回归进行改进，使其预测的结果更加精准，不失为对原始方法进行一种创新，也能够在原来的思想上带入一些新的思路与想法。

1.1.3 现实意义

我们不仅可以利用稀疏框架对Lasso的估计方法进行一系列的处理得到改进，同时伴随着机器学习的不断发展，深度学习的逐渐兴起，以及聊天机器人、Alpha Go、无人驾驶、无人机等一系列人工智能产物的诞生，人们的视线也逐渐被吸引过来。我们可以利用稀疏框架自身所具备的实用性和可塑性，不仅在理论上可以实现一定的进步，同时也能够与当前的技术相结合，因此也具备一定的实际意义。于此同时，Lasso作为一种改进的最小二乘估计，是最基本、最常用的特征选择的方法，具有连续缩小、泛化误差等优点，在建模和预测学习中具有很强的吸引力，在实际应用中非常有效。尤其是对于一个小n，大p的微矩阵时，Lasso方法能够很好地筛选出重要的特征，并且如果我们对于估计参数具有一定的先验信息，与Lasso方法相结合，我们既可以筛选不要的特征，又可以使我们需要的特征预测的更加准确，能对我们最终的筛选结果具有更高的一个预测精度，更小的泛化误差。综合上诉分析Lasso 回归模型在实际应用中有着十分巨大

的意义，因此，结合稀疏框架和先验信息的Lasso 同样具有相应的实际应用。

1.2 国内外文献综述

Lasso模型使用了L1范数作为惩罚项，这种惩罚项已广泛应用于统计学、机器学习、金融等多个领域。Lasso模型是1996年由Tibshirani(1996)提出来的，他是使用L1惩罚因子解决含有独立高斯分布的线性回归问题，虽然Lasso模型出现的比较晚。但是对于Lasso 模型的研究和探索一直在进行着高速的发展。也出现了很多改进后的模型，算法，并且应用于多个领域。

弹性网模型是Zou(2005)提出来的一种Lasso的变体，是岭回归和Lasso回归的一种折中模型，它的出现是为了解决Lasso不擅长处理的高度相关变量。融合Lasso 是由Tibshirani(2005) 提出来的，其最早是用于将基因的染色体进行降噪处理而产生的,现在不同的版本的融合Lasso已经有多种求解算法(Tibshirani等,2005,Hoefling等,2010,Tibshirani等,2011)。 Yuan(2006a,2006b,2006c) 引入了组Lasso，为了解决协变量本身具有组结构，他们发表的论文在当时引起了研究热潮。She(2010)提出了聚类Lasso，其在融合Lasso 的基础上更加进一步的对特征进行处理，试图使非零系数在聚类中相等，并且保持模型的稀疏性。其引用了稀疏框架为了更好的表示聚类Lasso。Negahban(2012) 提出了一个M 估计分析的一般性框架，其中组Lasso 只是一个特例。但是上述的模型都没有考虑过能否利用参数自身具备的某些先验信息，将其更好的发掘和利用，因此，本文重点研究的就是基于Lasso模型具有先验信息的情况下进行的一种改进和创新。

并且Lasso模型在多个领域中起着非常重要的作用。在解决稀疏规划问题时，他可以解释为寻找最小二乘或线性回归问题的稀疏问题，即采用若干个非零的变量。Lasso问题在信号处理领域中同样起着重要的作用，包括解决稀疏逆协方差(Yuan等,2012)、稀疏图回归(Zhou 等，2014) 和稀疏字典的学习(Sun等,2014)。Tibshirani(2005)将Lasso应用于生物数据的分析, 比如选取大数据中的一小部分来预测结果。于此同时，Lasso模型同样应用与视频分析中，Geng(2012)采用并行Lasso 的方法来解决大规模视频概念检测Zhu(2013)使用Group Lasso 解决视频标签以及Zhou(2013) 用稀疏离群值分割移动目标。在图像处理中，Afonso(2010)使用Lasso 方法解决图像修复问题，图像去噪(Elad 等,2006) 和去模糊口(Figueiredo等,2003)，其中正则化项是图像的梯度范数、网页图像排序(Yu 等,2014)及图像质量评估(He 等,2012)，利用稀疏编码处理图像标签问题(Dupe 等，2009)。在遥感领域中，Lasso 问题用来解决稀疏解(Bioucas-Dias等,2010,Plaza A,2011,Richards,1999)，从而可以得到每个端元都含有哪些物质，同

时在航天、农业领域，对研究物质成分问题起着非常重要的作用。

解决Lasso模型的方法也有很多，其中Tibshirani在原论文中提出一种可以解决Lasso问题的算法，但是该算法求解速度太慢。紧接着，Efron(2004) 提出了最小角回归，它能够有效地解决潜在的Lasso 优化问题，同时在统计和机器学习领域，它为加速Lasso算法提供了一个重要的工具。在解决大规模数据问题中，梯度下降方法是最简单的方法，Nesterov(2004,2005) 提出了最优的梯度下降方法和光滑模型，其在Lasso 模型中被广泛的应用。于此同时，由于L1 惩罚项的不可导，利用坐标下降法可以很好的解决这个问题，坐标下降法最早是Friedman(2007)发明的，而Wu(2008)则将其运用到Lasso 模型当中，Nesterov(2012) 利用坐标下降法改进了梯度下降法的不足。使得其在处理Lasso问题时，能够更好的适应大规模的数据，其中Saha(2013) 对于坐标下降法的收敛性质已经给出了详细的分析。近些年，Boyd(2010) 提出交替方向乘子(Alternating direction method of multipliers，ADMM)方法, 结合分布式凸优化问题，同样适合解决大规模数据的问题，并且已经成功地应用在很多领域，包括解决Lasso模型的求解问题。于此同时，Goldstein T(2014) 提出了快速ADMM方法。最近的研究者Suzuki(2014)结合大数据的特点采用随机的方法，在原有算法的基础上，提高了算法的收敛效率。

通过梳理上诉求解Lasso模型的方法后，本文需要求证当前的主流求解方法是否适用于本论文改进后Lasso模型。如果不适用需要在原有的求解方法基础上进行改进，并且对各个方法的收敛时间和准确度进行一定的对比。

1.3 本文结构

第1章是绪论部分，在本章的开始介绍了Lasso估计问题和在稀疏框架下具有先验信息方法的研究背景及研究意义，在研究意义中包含了理论意义和现实意义。接下来介绍了国内外在此方面的研究状况。随后对相关的背景知识做了简单的介绍，包括最小二乘估计方法和岭回归。

第2章全面阐述了Lasso回归模型，首先对于Lasso回归模型的背景做了介绍，紧接着给出了Lasso回归模型的定义，并且简要证明了两种不同形式下的Lasso回归模型具有一一对应的关系。在这之后给出了Lasso回归的三种求解方法：Lasso算法、最小角回归算法和坐标下降法，与此同时给出了每种算法的具体步骤和优缺点。最后，本节针对Lasso 回归模型的性质进行了归纳总结，并且给出了相应的证明。

第3章研究了本文的理论研究部分，首先本章引入了一种更为宽泛的Lasso 回归模型，即基于稀疏框架的Lasso回归，并在此之上融入了先验信息的思想。紧接着给出了新模型的具体形式，并且就两种不同形式的T给出了相应的求解算法，最后说明了新模

型所具有的良好性质，为后文的数值实验做出理论支撑。

第4章是对新模型进行数据实验，本章将传统Lasso 模型和新模型在模拟和实例数据上分别进行研究。具体地，在模拟数据实验中我们计算并比较各个模型的均方误差和预测误差，并且画图描述了各模型计算的参数值与真实参数的趋近状况。在实例数据分析中我们计算不同模型的平均预测误差和平均绝对误差，表现其在真实数据中的可行性。

第5章是总结和展望，总结本文所研究的主要问题、方法与结果，并描述未来的扩展性工作与研究方向。

1.4 课题相关背景知识介绍

1.4.1 最小二乘估计

在线性回归模型中的回归系数的估计问题一直是众多学者研究分析的课题，其中最原始也是最基本的估计方法就是最小二乘估计，它通过最小化误差的平方和寻找数据的最佳函数匹配，已达到最终解决问题的目的。

首先给出一个线性模型。其中存在着m 个自变量$X_1, X_2, \cdots, X_m$ 和因变量y, 具有如下的线性关系

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m + \varepsilon. \tag{1-4}$$

其中$\beta_0, \beta_1, \beta_2, \cdots, \beta_m$为未知参数，也是线性模型中要估计的对象，$\varepsilon$是期望为0 方差为$\delta^2$ 的误差项.

若对于此线性模型有n组观测值，即$(x_{i1}, x_{i2}, \cdots, x_{im}, y_i)$,i=0,1,$\cdots$,n, 那么误差项$\varepsilon = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)$,并且有$Cov(\varepsilon_i, \varepsilon_j) = \begin{cases} 0 & i \neq j \\ \sigma^2 & i = j \end{cases}$，则第i组观测值的线性关系为

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im} + \varepsilon_i, i = 1, 2, \cdots, n. \tag{1-5}$$

其中矩阵形式为

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1m} \\ 1 & x_{21} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}. \tag{1-6}$$

将此方程简化为

$$y = X\beta + \varepsilon. \tag{1-7}$$

最小二乘估计的原理为使得公式1-1达到最小值，通过求偏导得到参数估计

$$\hat{\beta}^{OLS} = (X'X)^{-1}X'y. \tag{1-8}$$

使用最小二乘估计后的参数估计具有如下性质：

性质1.1 β的估计$\hat{\beta}$是β的无偏估计.

性质1.2 若$\varepsilon \sim N(0,\sigma^2 I)$,则有$\hat{\beta} \sim N(\beta,\sigma^2(X'X)^{-1})$.

性质1.3 假设b为$(m+1)\times$的常数向量，那么对于线性函数$b'\beta$, 称$b'\hat{\beta}$ 为$b'\beta$的最小二乘估计，那么有$(b'\hat{\beta} \sim N(b'\beta,\sigma^2 b'(X'X)^{-1}b)$

性质1.4 残差平方和为RSS，则有$\hat{\sigma^2} = \frac{RSS}{n-m-1}$ 是σ^2 的无偏估计

1.4.2 岭回归估计

上一节说到最小二乘估计的原理是使得$\|y - X\beta\|^2$ 达到最小值，我们将其作为度量，引入均方误差(Mean Squared Error) 来衡量参数估计量接近真实参数的程度

$$MSE(\hat{\beta}) = E(\left\|\hat{\beta} - \beta\right\|^2) = Var(\hat{\beta}) + [E(\hat{\beta}) - \beta]^2. \tag{1-9}$$

因此，为了使均方误差最小化，可以视为在估计量方差与偏差之间进行权衡。于此同时我们引入预测误差(Predicted Error)的定义

$$PE = E(\|\hat{y} - y\|^2) = MSE + \sigma^2. \tag{1-10}$$

由上式可知，在给定σ^2的前提下，MSE最小化等价于PE最小化。在多元线性回归模型中，用最小二乘估计得到的回归系数估计量是具有无偏性的且具有最小方差。但是并不能说明最小二乘估计量的方差一定是最小的。尤其是在模型存在多重共线性的问题时候，一个稍稍有偏的估计量，其精度可能远远高于无偏估计量。于是提出了一种以放弃对线性回归模型中的回归系数一般最小二乘估计的无偏性要求的方法，也就是岭回归估计的诞生。

在多元线性回归模型中，回归方程出现多重共线性时，会存在$\left|X'X\right| = 0$ 的情况，这样的情况会使得参数估计

$$\hat{\beta} = (X'X)^{-1}X'y. \tag{1-11}$$

变得很不稳定。并且会导致$\hat{\beta}$的均方误差趋向于无穷大。这时，如果在$(X'X)$ 中加入一个常数矩阵$kI(k>0)$,那么我们就可以得到$(X'X+kI)$ 形成的矩阵一定是非奇异的。

从而得到岭回归估计(Arthur等,1970)的参数表达式为

$$\hat{\beta}^k=(X'X+kI)^{-1}X'y. \tag{1-12}$$

其中的参数k被称作岭参数。因此，线性回归方程的目标函数则由原来的公式1-1 转变为

$$min\frac{1}{2}\|y-X\beta\|_2^2+k\|\beta\|_2^2. \tag{1-13}$$

岭回归估计具有以下几点性质：

性质1.5 当岭参数k被认为是和y 无关的参数时，岭回归估计的$\hat{\beta}^k$ 是最小二乘估计$\hat{\beta}$的线性变换。

性质1.6 $((\hat{\beta}_k))=A_k\beta$,其中$A_k=(X'X+kI)^{-1}X'X$, 只要$A_k\neq I$,岭回归估计的$\hat{\beta}^k$ 就是β的有偏估计

性质1.7 岭回归估计$\hat{\beta}_k$是$\hat{\beta}$向原点的压缩，即对任意$k>0$, 且$\left\|\hat{\beta}^k\right\|\neq 0$ 的情况，总有$\left\|\hat{\beta}^k\right\|<\|\beta\|$.

性质1.8 存在$k>0$，在均方误差准则下，岭回归估计要优于最小二乘估计。

1.5 本章小结

这一章在介绍了本文研究的课题背景以及研究课题的理论意义和实际意义之外，还介绍了国内外对此的研究现状。介绍完相应的内容之后对本章的整体结构做了一个简单的介绍。最后，对于本文需要的相关背景知识也进行了简要说明。通过介绍最小二乘估计和岭回归估计这两种参数估计的方法，引出我们后续需要研究的内容。

第 2 章　Lasso回归模型

2.1　Lasso回归的背景

通过绪论的相关知识介绍，我们了解到对于一般线性回归方程，$Y = X\beta + \varepsilon$，最小二乘估计得出的估计量$\hat{\beta}_{OLS}$具有以下优点:(1) 方法简单易行，计算并不复杂.(2)在参数估计量无偏的基础上均方误差最小。于此同时，缺点很明显：(1) 如果矩阵X 不是列满秩，则$\hat{\beta}_{OLS}$ 结果不唯一;(2) 当自变量之间存在高度相关性时，会导致$(X^{'}X)^{-1}$增大，从而导致$MSE(\hat{\beta})$增大。

面对这种情况，学者们提出了惩罚函数，整体思路是通过牺牲一定的无偏性，进而降低估计量整体的均方误差，进而提高模型整体的预测误差。其中典型的代表就是岭回归估计，$\hat{\beta}_k = argmin\frac{1}{2}\|y - X\beta\|_2^2 + k\|\beta\|_2^2$，岭回归的建立有效的解决了最小二乘估计中当自变量存在高度相关时导致$MSE(\hat{\beta})$增大的情况。并且在$k > 0$时，岭回归估计结果一般都优于最小二乘估计。尽管它的预测准确性较好，但在大多数情况下，它不能将回归系数收缩到0，导致回归方程中的变量数目仍然较多，这不利于我们我们找到和预测结果相关性最高的变量。

随后，为了解决这个问题，Breiman(1995)年提出了non-negative garrotte(简称为NNG)，其定义如下

$$\hat{\beta}^{NNG} = agrmin\sum_{i=1}^{N}(y_i - \sum_{j=1}^{p} c_j\hat{\beta}_j^{ols}x_{ij})^2 + \lambda\sum_{j=1}^{p}\left|c_j\right|. \tag{2-1}$$

其中$\lambda > 0$，随着λ的不断增大，c_j中部分元素被压缩至0，从而得出$\hat{\beta}_j^{nng} = c_j\hat{\beta}_j^{ols}$。在原论文中，由于初始$\hat{\beta}^{NNG}$是通过普通最小二乘估计方法得到的，因此最小二乘估计存在的缺点在NNG中依然存在。当自变量个数大于样本数时，由于$\hat{\beta}^{OLS}$ 不稳定，故$\hat{\beta}^{NNG}$也不稳定.但是后来其他研究人员证明：用其他初始估计(如Lasso、岭回归、弹性网)时，NNG具有非常好的性质。

NNG的诞生与本文研究的重点Lasso关系紧密，正是Breiman的论文给1996 年的Tibshirani 的论文提供了Lasso 灵感。

2.2　Lasso回归的定义

假定我们有数据$(x^i, y_i), i = 1, 2, \cdots, N$,其中$x^i = (x_{i1}, \cdots, x_{ip})^T$ 作为解释变量，y_i作为

响应变量。对于一般的线性回归而言，我们假定观测对象是独立的，并且解释变量$\boldsymbol{x}^i$已经进行归一化，即$\frac{1}{N}\sum_{i=1}^{N} x_{ij}=0, \frac{1}{N}\sum_{i=1}^{N} x_{ij}^2=1$。如果不进行归一化，则Lasso得到的结果会受到具体测量单位的影响。并且为了方便起见，我们假定响应变量y_i已经进行中心化，即$\frac{1}{N}\sum_{i=1}^{N} y_i=0$。

令$\hat{\beta}=(\hat{\beta}_i,\cdots,\hat{\beta}_p)^T$,则Lasso 的定义为

$$\hat{\beta}^L = argmin\frac{1}{2}\sum_{i=1}^{N}(y_i-\beta_o-\sum_{j=1}^{p}x_{ij}\beta_j)^2, s.t.\sum_{j=1}^{p}\left|\beta_j\right|\leq t. \tag{2-2}$$

其中$t(t\geq 0)$称为调和参数，调和参数t的大小控制着回归系数总体的大小。若令$t_0=\sum_{j=1}^{p}\left|\beta_j\right|$, 则$t\leq t_0$。就会使一些回归系数缩小并最终趋向于0，一些较小的系数甚至直接缩小至0。t值越大，各个系数的取值范围就越大，模型从而更有能力拟合训练数据。反之，t值越小，系数的取值范围越小，模型从而更加稀疏，有更好的可解释性。因此，需要找到合适的t 值，在这两种情况下找到一种平衡。

由于先前我们假定y_i已经进行中心化，因此可在Lasso优化中省略截距β_0。如果y_i未进行中心化处理，则非中心化数据得到的截距β_0 可通过下面的公式来计算

$$\hat{\beta}_0=\bar{y}-\sum_{j=1}^{p}\bar{x_j}\hat{\beta}_j. \tag{2-3}$$

其中$\bar{y}$和$\bar{x_j}$表示样本均值。因此，本文会忽略Lasso 的截距β_0。因此公式2-2转化为

$$\hat{\beta}^L = argmin\frac{1}{2}\sum_{i=1}^{N}(y_i-\sum_{j=1}^{p}x_{ij}\beta_j)^2, s.t.\sum_{j=1}^{p}\left|\beta_j\right|\leq t. \tag{2-4}$$

同时，为了方便后续讨论，公式2-4一般改写成拉格朗日形式，如下所示

$$\hat{\beta}^L = argmin\frac{1}{2}\left\|y-X\beta\right\|_2^2+\lambda\left\|\beta\right\|_1. \tag{2-5}$$

其中$\lambda\geq 0$。上式之所以被称为拉格朗日形式，是因为根据拉格朗日对偶性质可知，λ与t 是一一对应的关系。下面简单的证明当$p=2$ 时，且X是正交矩阵时，λ 与t一一对应。不过在此之前引入软阈值函数的概念，其定义如下

$$\eta_s(w,\lambda)=sgn(w)(|w|-\lambda)_+. \tag{2-6}$$

其中,sgn为符号函数，当$|w|>\lambda$时，$(|w|-\lambda)_+=|w|-\lambda$,当$|w|\leq\lambda$ 时,$(|w|-\lambda)_+=0$,因

此软阈值函数等价于

$$\eta_s(w,\lambda)=\begin{cases} w+\lambda, & w<-\lambda \\ 0, & |w|\le\lambda \\ w-\lambda, & w>\lambda \end{cases}. \tag{2-7}$$

在满足KKT的条件下，$\hat{\beta}$ 是最优解当且仅当

$$X^T(X\hat{\beta}-y)+\lambda s=0. \tag{2-8}$$

其中s表示$\|\beta\|_1$的次梯度。对于绝对值函数，其次梯度的形式为$s\in sgn(\beta)$,即当$\beta\neq 0$时，$s=sgn(\beta)$;$\beta=0$,$s\in[-1,+1]$。当X是正交矩阵时，$X^TX=E$, 则$\hat{\beta}^L=X^Ty-\lambda sgn(\beta)$ 显然，λ和t构成以下方程组

$$\begin{gathered} \hat{\beta}_1^L=sgn(\hat{\beta}_1^{OLS})(\left|\hat{\beta}_1^{OLS}\right|-\lambda)_+, \\ \Sigma_1^p\left|\hat{\beta}_j^L\right|=t. \end{gathered} \tag{2-9}$$

不失一般性，假设$\hat{\beta}_j^L$均为正数，则如下方程组

$$\begin{aligned} \hat{\beta}_1^L&=(\hat{\beta}_1^{OLS}-\lambda)_+, \\ \hat{\beta}_2^L&=(\hat{\beta}_2^{OLS}-\lambda)_+, \\ \hat{\beta}_1^L+\hat{\beta}_2^L&=t. \end{aligned} \tag{2-10}$$

解上式可得

$$\lambda=\frac{\hat{\beta}_1^{OLS}+\hat{\beta}_2^{OLS}-t}{2}. \tag{2-11}$$

证毕。

衍生至一般情况

$$t=\sum_{i=1}^{p}sgn(\hat{\beta}_j{}^0)\hat{\beta}_j{}^0-p\lambda. \tag{2-12}$$

通过上述可知当λ越大，则t越小，即系数越趋近于0,模型从而更加稀疏。反之λ越小，则t越大，即系数越不趋近于0，模型从而不呈现稀疏性。

2.3 Lasso回归的求解

从Lasso的定义便可以看出，求解$\hat{\beta}^L$估计量是一个凸优化问题，实质上是解一个带不等式约束的二次规划问题(Quadratic program,QP)。因此，有许多QP 方法可以用来求

解Lasso。Tibshirani(1996) 的文章中给出了一个Lasso 算法，但是由于Lasso 算法中存在着很大的限制。Efron(2004)提出了一个最小角回归算法，该算法会将整个解的路劲作为惩罚参数λ的一个函数，但它并不适合大规模问题，进而Friedman(2007)和Wu(2008) 首先将坐标下降法运用到Lasso 模型中，其在解决高维数据和大规模样本中起着至关重要的作用。本节将简要介绍上诉算法的思想和步骤。

2.3.1 Lasso算法

首先固定$t \geq 0$,公式(2.4)可以表现为具有2^p个不等式约束的最小二乘估计，简单来说就是对应于p个系数的2^p种组合。Lawson、Hansen(1974) 提出了解满足线性不等式$G\beta \leq h$ 的最小二乘估计的方法。但是该方法在Lasso 中难以直接运用，但是可以通过逐步加入约束条件来解决，最终找到一个可行的，满足KKT条件的解，具体算法如下：

1. 令$\delta_i = sgn(\hat{\beta})$，这里的$\hat{\beta}$ 通常是$\hat{\beta}^{OLS}$ 估计量。

2. 计算$\hat{\beta} = argmin\sum_{i=1}^{N}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2, s.t. G\beta \leq t$，这里$G = \delta_i^T$。

3. 验证是否满足$\sum_{j=1}^{p} \left|\beta_j\right| \leq t$, 若满足则停止，$\hat{\beta}$即为解，否则继续。

4. 令$\delta_j = sgn(\hat{\beta})$ 加入到G中，即令$G = \begin{pmatrix} \delta_i^T \\ \delta_j^T \end{pmatrix}$，然后返回2步。

这个算法优点在于逐步加入约束条件，使计算较为简单。但是缺点也很明显：(1) 当自变量较多时，该算法的速度较慢，对于大数据并不适用。（2）当X 列向量不满秩时，该算法不适用。

2.3.2 最小角回归算法

概括来说最小角回归算法(Least Angle Regression,简称LARS)与向前选择法(Forward Selection) 类似。向前选择法是通过每次增加一个变量来建立模型，即每步都会挑出最好的变量放到模型中，然后用模型中所有的变量来进行最小二乘拟合。而最小角回归采用了相同的策略，但只会尽可能多地输入适合模型的变量。第一步得到与输出最相关的变量。LARS会不断朝最小二乘值移动该变量的系数（这会让它与残差的相关性在绝对值上减少），而不是拟合该变量。一旦别的变量与残差的相关性达到该值，这个过程就会暂停，并将第二个变量加到活动集中，按相同方式修改它们的系数，并将相关系数绑定，共同减少。当模型中的所有变量都进行了一次最小二乘拟合后，这个过程会停止。具体算法如下：

1. 首先对训练样本进行归一化，使其样本均值为0。初始残差为$r_0 = y - \bar{y}, \beta^0 = (\beta_1, \beta_2, \cdots, \beta_p) = 0$。

2. 找到与r_0最相关的x_j，即$\rho_0 = max\left|\langle x_j, r_0\rangle\right|$, 定义$A = \{j\}$和一个由单变量构成的矩阵$X_A$。

3. 对于$k = 1, 2, \cdots, K = min(N-1, p)$, 进行以下操作：

   (a) 定义最小二乘方向$\delta = \frac{1}{\rho_{k-1}}(X_A^T X_A)^{-1} X_A^T r_{k-1}$，对于p 维向量$\Delta$, 使$\Delta_A = \delta$，其他元素都为0。

   (b) 以Δ方向，从β^{k-1}开始，朝着它们的最小二乘解移动系数β,其中$X_A : \beta(\rho) = \beta^{k-1} + (\rho_{k-1} - \rho)\Delta$, $0 \le \rho \le \rho_{k-1}$。进一步得到新的残差$r(\rho) = y - X\beta(\rho) = r_{k-1} - (\rho_{k-1} - \rho)X\Delta$。

   (c) 关注$\rho = |\langle x_l, r(\rho)\rangle|$,设当$l$个变量让$\rho$ 达到最大，则应该将l 加入A，且令$\rho_k = \rho$。

   (d) 让$A = A \cup \{l\}$,$\beta^k = \beta(\rho_k) = \beta^{k-1} + (\rho_{k-1} - \rho_k)\Delta$,$r_k = y - X\beta^k$。

4. 返回序列$\{\rho_k, \beta^k\}_0^K$

解释下该算法第三步中的K，如果$p > N-1$，在经过$N-1$步之后，LAR会得到残差为0的解。事实上，从算法步骤来看LAR在运算上是节俭的，其计算的消耗与通常的最小二乘估计是一样的。但是对于解释变量是$n \times p$的情形，该方法最多选的变量个数为$min(N-1, p)$,用该算法往往会得到过于稀疏的模型。

2.3.3 坐标下降法

Wu(2008)首次提出坐标下降法在Lasso中的应用，其本质是一种迭代算法，在单个坐标方向执行β^t到β^{t+1}的迭代，然后在该坐标方向上求单变量最小值。更准确的讲，如果第t 次迭代选择了坐标k。具体的算法过程如下：

1. 首先对于β随机取一个初值，记为β^0，表示初始轮数。

2. 对于第t轮的迭代。我们从β_1^t开始，到β_p^t 为止，依次求β_k^t。具体表达式为：$\beta_k^{t+1} = \underset{\beta_k}{argmin} f(\beta_1^t, \cdots, \beta_{k-1}^t, \beta_k^t, \beta_{k+1}^t, \cdots, \beta_p^t)$。其中$\beta_j^{t+1} = \beta_j^t, j \ne k$,因此最小值很容易通过求导求得。

3. 当$\left|\beta^k - \beta^{k-1}\right|_1$ 降低到规定阈值，迭代结束，否者重复2步，继续$k+1$次迭代。

在Lasso回归求解中，当N非常大时，求全梯度或者次梯度需要消耗大量的时间和储存空间，而坐标下降法可以通过求解目标函数的子函数来求解最优解。这里的子函数可以看做是坐标系中的某一个坐标或某一组坐标，保持其他的坐标系的变量不变，每次迭代只优化步骤2 中的目标函数，因此和前两种算法比较，它的运行速度非常迅速，并且坐标下降法已经被广泛应用于Lasso 处理大规模数据中。

2.4 Lasso回归的性质

在过去十几年中，学者们已经得到了Lasso估计的诸多良好性质，本节对这些性质进行了归纳总结。

2.4.1 拟合值的唯一性

Tibshirani(2013)的文章中指出，如果X中的数据取自一个连续概率分布，则对于$\lambda > 0$，Lasso问题（公式2.5）的拟合值是唯一的。即使$p \geq N$，这一点任然成立，尽管任何Lasso 解的非零系数个数最多为N。

对于最小二乘估计而言，若矩阵X不是列满秩，其拟合值唯一，但参数估计本身不唯一。非满秩情形在$p \leq N$时可能出现共线性问题，在$p > N$时则总会出现。对于后一种情况,$\hat{\beta}$的解有无限多个，这些解都会让训练误差为零。现在考虑在$\lambda > 0$的情形下，假设有两个Lasso解$\hat{\beta}$ 和$\hat{\gamma}$,对应的最优解的值为c^*,则有$\left\|\hat{\beta}\right\|_1 = \|\hat{\gamma}\|_1$。下面进行证明。

（1）假设$\hat{\beta}$和$\hat{\gamma}$为Lasso模型的解，且$\hat{\beta} \neq \hat{\gamma}$。因为二者都是该问题的最优解，因此我们知道$\alpha\hat{\beta} + (1-\alpha)\hat{\gamma}$,对于任意$0 < \alpha < 1$ 而言，都为该问题的最优解。

（2）已知$\hat{\beta} \neq \hat{\gamma}$，则$X\hat{\beta} \neq X\hat{\gamma}$。对于任何$0 < \alpha < 1$，有以下不等式

$$\frac{1}{2}\left\|y - X(\alpha\hat{\beta} + (1-\alpha)\hat{\gamma})\right\|_2^2 + \lambda\left\|\alpha\hat{\beta} + (1-\alpha)\hat{\gamma}\right\|_1 < \alpha c^* + (1-\alpha)c^* = c^*. \tag{2-13}$$

这意味着$\alpha\hat{\beta} + (1-\alpha)\hat{\gamma}$获得了一个比$c^*$ 更小的值，这明显是矛盾的。

（3）通过2步，我们知道$X\hat{\beta} = X\hat{\gamma}$必须成立，才能得到相同的均方误差，因此根据$l_1$惩罚性的凸性，则对于任何$\lambda > 0, \left\|\hat{\beta}\right\|_1 = \|\hat{\gamma}\|_1$。详细的证明在Tibshirani的论文中给出了这方面的总结。不过计算Lasso的数值算法通常与算法的具体细节有关。采取不同的算法，初始值的选择会影响最终的解。

2.4.2 估计量的相合性

在绪论中本文介绍了最小二乘估计和岭回归估计的一致性，那么在Lasso中需要什么附加条件可确保参数向量差值$\Delta\beta = \left\|\hat{\beta} - \beta\right\|_2$也收敛于0。Knight 和FU(2000) 通过研究证明了在自变量个数p 和真实估计量β 不变的情况下，Lasso估计量具有一致性。下面对其证明：

首先我们做出如下假设：(1)$\Sigma^n = \frac{1}{n}X_n^T X_n \to \Sigma$，且$\Sigma^n$非奇异或者列满秩。(2)$\frac{1}{n}\underset{1\le i\le n}{max}\, x_i^T x_i \to 0$。(3)$\lambda_n = o(n)$，即$\lambda_n/n \to 0$,且$\lambda_n$表示会随着样本量$n$ 的变化而变化。

在上述条件中，当x是i.i.d且是二阶矩有限，则$\Sigma = E((x_i^n)^T x_{i^n}), \frac{1}{n}X_n^T X_n \to \Sigma, \underset{1\le i\le n}{max}\, x_i^T x_i = o_p(n)$, 因此前两个条件很好满足。前文已经说过为了确保没有量纲的影响，已经对X 进行归一化，对Y 进行中心化。则

$$\hat{\beta} = argmin\frac{1}{2}\left\|y - X\beta\right\|_2^2 + \lambda_n\left\|\beta\right\|_1. \tag{2-14}$$

并且要想上式存在最优解，在满足KKT 的条件下，$\hat{\beta}$是最优解当且仅当

$$X^T(X\hat{\beta} - y) + \lambda_n s(\hat{\beta}) = 0. \tag{2-15}$$

将$\hat{\beta}$提到方程左边，则

$$\begin{aligned}\hat{\beta} &= \frac{1}{n}\Sigma^{-1}(X^T y - \lambda_n s(\hat{\beta}))\\ &= \frac{1}{n}\Sigma^{-1}(X^T(X\beta + \varepsilon) - \lambda_n s(\hat{\beta}))\\ &= \beta + \frac{1}{n}\Sigma^{-1}(X^T\varepsilon - \lambda_n s(\hat{\beta})).\end{aligned} \tag{2-16}$$

当$\lambda_n = 0$时,$(\hat{\beta} - \beta) = \frac{1}{n}\Sigma^{-1}X^T\varepsilon \sim N(0, \frac{\Sigma^{-1}}{n}\sigma^2) = O(\frac{1}{\sqrt{n}}) = o(1)$, 很明显，$\hat{\beta} \xrightarrow{p} \beta$。这就是最小二乘估计的无偏性。当$\lambda_n \neq 0$时

$$\begin{aligned}(\hat{\beta} - \beta) &= \frac{1}{n}\Sigma^{-1}(X^T\varepsilon - \lambda_n s(\hat{\beta})\\ &= \frac{1}{n}\Sigma^{-1}X^T\varepsilon - \frac{\lambda_n}{n}\Sigma^{-1}s(\hat{\beta}).\end{aligned} \tag{2-17}$$

公式前半部分就是最小二乘估计的无偏性，而公式的后半部分，由于先前假设过$\lambda_n = o(n)$，因此$\frac{\lambda_n}{n}\Sigma^{-1}s(\hat{\beta}) = \frac{\lambda_n}{n}O(1) = o(1)$。证毕。因此Lasso估计量在上述条件成立下是一致估计量。

2.4.3 变量选择一致性

对于Lasso估计$\hat{\beta}$是否如真实回归向量β一样在相同的位置有非零项。具体而言，假设真实回归向量β是k稀疏的，目标是正确的找出与真实回归相同的具有相同集合的最优解$\hat{\beta}$。这个性质称为变量选择一致性，也可以叫做稀疏性。其定义如下

$$P(\left\{i:\hat{\beta}_i^n \neq 0\right\} = \{i:\beta_i^n \neq 0\}) \to 1, as \quad n \to \infty. \tag{2-18}$$

即当样本量$n \to \infty$时，通过该方法选择出正确自变量的概率趋近于1。在变量选择一致性的基础上，Zhao和Yu（2006）提出了符号一致性，其定义如下

$$P(sgn(\hat{\beta}_i) = sgn(\beta_i)) \to 1, as \quad n \to \infty. \tag{2-19}$$

其中，sgn是符号函数，显然符号一致性是强于变量选择一致性，前者仅要求非零系数的位置一致，而后者要求非零系数的位置和符号均一致。于此同时，作者在符号一致性的基础上，根据正则化的强弱定义了两种符号一致性。

强符号一致性：当存在$\lambda_n = f(n)$，且函数独立Y 和X。

$$\lim_{n\to\infty} P(sgn(\hat{\beta}^n(\lambda_n)) = sgn(\beta^n)) = 1. \tag{2-20}$$

弱符号一致性：存在$\lambda \geq 0$。

$$\lim_{n\to\infty} P(sgn(\hat{\beta}^n(\lambda)) = sgn(\beta^n)) = 1. \tag{2-21}$$

强符号一致性意味着有可以使用预先选好的λ来实现Lasso的变量选择一致性。而弱符号一致性则表示对于随机的正则化参数需要存在一个正确的λ 来选出真实的模型。很明显强符号一致性$\supseteq$弱者符号一致性的。如何才能达到上述的一致性结果，作者在原文中给出了相应条件和要求。

假定$\beta^n = [\beta_1^n, \beta_2^n]^T$，$\beta_1^n = 0; \beta_2^n \neq 0$，且$\beta_1^n$ 和β_2^n的规模为$d_1 \times 1, d_2 \times 1, d_1 + d_2 = p$。令

$$\Sigma^n = \frac{1}{n} X_n^T X_n = \frac{1}{n} \begin{bmatrix} X_1^T X_1 & X_1^T X_2 \\ X_2^T X_1 & X_2^T X_2 \end{bmatrix} = \begin{bmatrix} \Sigma_{11}^n & \Sigma_{12}^n \\ \Sigma_{21}^n & \Sigma_{22}^n. \end{bmatrix} \tag{2-22}$$

在Σ_{11}^n可逆的前提下，存在一个正向量η，作者提出了**强不可代表条件**

$$\left|\Sigma_{21}^n(\Sigma_{11}^n)^{-1}sgn(\beta_1^n)\right| \leq 1-\eta. \tag{2-23}$$

和**弱不可代表条件**

$$\left|\Sigma_{21}^n(\Sigma_{11}^n)^{-1}sgn(\beta_1^n)\right| < 1. \tag{2-24}$$

两者的不同点在于左式接近与1时，弱不可代表条件恒成立，而强不可代表条件在极限处不存在。接下来本节分别讨论在常规数据和高维数据情况下的变量选择一致性。

对于固定p且$\beta^n=\beta$,常规数据下

1. 假定$\Sigma^n=\frac{1}{n}X_n^TX_n\to\Sigma$，其中$\Sigma$是正定矩阵。
2. $\frac{1}{n}\underset{1\leq i\leq n}{max}((x_i^n)^Tx_i^n)\to 0$。
3. 强不可表示条件。

当上述条件只有前两条成立时，当且仅当存在N使得弱不可代表条件成立后，Lasso模型符合弱符号一致性；当强不可代表也成立时，Lasso模型符合强符合一致性，对于$\forall\lambda_n$符合$\frac{\lambda_n}{n}\to 0,\frac{\lambda_n}{n}^{(\frac{1+c}{2})}\to\infty$，且$0<c<1$, 则

$$P(sgn(\hat{\beta}^n(\lambda_n))=sgn(\beta^n))=1-o(e^{nc}). \tag{2-25}$$

因此强不可表示条件成立，那么Lasso 选择真实模型的概率以指数接近于1，且从上一小节可知对于$\lambda_n=o(n)$时，Lasso也具有一致估计量，因此强不可条件允许同时进行一致的模型选择和参数估计。另一方面，即使对于弱符号一致性，弱不可代表条件也是必要的。因此，在常规数据下，强不可代表条件$\supseteq$强符号一致性$\supseteq$弱符号一致性$\supseteq$弱不可代表性。

高维数据情况下，由于d_1,d_2 会随着n的增长而增长，因此令$d_1=d_1^n,d_2=d_2^n$, 并且Σ^n和β^n也会随着n 的增长而变化。假定存在$0\leq c_1\leq c_2\leq 1,M_1,M_2,M_3,M_4>0,K>0$ 使得

1. $\frac{1}{n}(X_i^n)^{-1}X_i^n\leq M_1$。
2. $\alpha^{-1}\Sigma_{11}^n\alpha\geq M_2,\forall\|\alpha\|_2^2=1$。
3. $d_1^n=O(n^{c_1})$
4. $n^{\frac{1-c_2}{2}}\underset{i=1,\cdots,d_1}{min}|\beta_i^n|\geq M_3$。

5. $E(\varepsilon_i^n)^{2k} < \infty$。

6. 强不可表示条件。

当上述条件均成立，且$\forall\lambda_n$满足$\frac{\lambda_n}{\sqrt{n}} = o(n^{\frac{c_2-c_1}{2}})$;$\frac{1}{d_1^n}(\frac{\lambda_n}{\sqrt{n}})^{2k} \to \infty$时。有

$$P(sgn(\hat{\beta}^n(\lambda_n)) = sgn(\beta^n)) = 1 - O(\frac{d_1^n n^k}{\lambda_n^{2k}}) \to 1 \; as \; n \to \infty. \tag{2-26}$$

因此对于高维数据而言，想要达到符号一致性不仅仅需要强不可表示条件，还需要对噪声进行相应的限制。

2.5 本章小结

本章主要对Lasso回归模型进行了一个详细的说明与概括。首先对于Lasso回归模型的背景进行了简要的介绍，说明其建立的原因和过程。之后给出了Lasso回归的定义，并且就Lasso的普通形式和拉格朗日形式一一对应的关系进行了简单的证明，于此同时引入了软阈值函数的概念。紧接着罗列了Lasso求解算法的具体思想和步骤，其中包括Lasso 算法、最小角回归算法和坐标下降法。最后，对于Lasso回归模型中具有的良好性质进行了相应的证明和概述，其中包括拟合值的唯一性，估计量的相合性和变量选择一致性（稀疏性）。

第3章　基于先验稀疏框架的Lasso 回归的理论研究

本章将在Lasso回归模型的基础上引入一个更一般的回归模型，即基于稀疏框架的Lasso 回归模型，并在此之后提出本文的核心内容，即加入先验信息后的Lasso回归模型。因此本文将首先在本章第一节给出相关的知识概念，其次在第二节介绍了新模型的定义，及其使用场景。最后在第三节给出了新模型的求解方法。

3.1　相关知识工作

3.1.1　稀疏框架

在过去几年中，稀疏性已经成为应用数学、计算机科学和电气工程各个领域的一个重要概念。在稀疏信号处理中，为了实现许多类型的信号只用几个非零系数表示，Robert等(2011) 提出了稀疏框架的概念，并进行了深入研究。稀疏框架的概念如下。

设$(e_j)_{j=1}^n$是空间$\mathbb{R}^n$上的一组标准正交基。如果对每个$i \in \{1,2,\cdots,m\}$，都存在$J_i \subseteq \{1,2,\cdots,n\}$，使得$\mathbb{R}^n$ 中的框架$F = \{f_i\}_{i=1}^m$中的每个元素都满足

$$f_i \in span\{e_j : j \in J_i\}. \tag{3-1}$$

和

$$\sum_{i=1}^{m} |J_i| = k. \tag{3-2}$$

那么称框架$F = \{f_i\}_{i=1}^m$ 是k阶稀疏框架，其中$|J_i|$ 表示J_i 中元素的个数，并记

$$d = |J|,\ J = \bigcup_{i=1}^{m} J_i. \tag{3-3}$$

显而易见，$1 \le d \le n$。令$F = \{f_i\}_{i=1}^m$，$e = (e_j)_{j\in J}$ 与$R = (r_{ij}) \in \mathbb{R}^{m\times d}$，则有

$$f_i = \sum_{j\in J_i} r_{ij} e_j,\ i = 1,2,\cdots,m. \tag{3-4}$$

即稀疏框架可以表示为

$$F = Re. \tag{3-5}$$

其中,R也被称为基系数，e为标准正交基构成的矩阵

3.1.2 基于稀疏框架的Lasso回归

在本节中，首先我们回顾一般的回归问题

$$y = X\beta + \varepsilon \; \varepsilon \sim N(0, \sigma^2 I). \tag{3-6}$$

其中y是因变量，X是$n \times p$的自变量矩阵。为了得到稀疏的β, 和更加精准的预测误差。Lasso是当前最流行、最基本的方法之一。它的拉格朗日形式为

$$\hat{\beta}^L = argmin \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 . \tag{3-7}$$

或以多目标的方式表示

$$min(\|y - X\beta\|_2^2, \|\beta\|_1). \tag{3-8}$$

接下来，我们引入上一节介绍的稀疏框架，将正则化中的β看成稀疏框架中的R，加入一个T 矩阵来代替原先的e，但区别在于，这里的T不在是由标准正交基构成的，而是由一个用户指定的稀疏矩阵来替代。可以看成是在Lasso模型基础上更为一般的形式，其定义为

$$min(\|y - X\beta\|_2^2, \|T\beta\|_1). \tag{3-9}$$

其中，T是任意构造的稀疏矩阵，其作用是通过稀疏编码来限定真实β 之间的关系。下面举出几个例子说明。

Example 3.1.1 (Lasso) 很明显，当T为单位矩阵时，原式就变为经典的*Lasso* 模型。

Example 3.1.2 (Fused Lasso) *Tibshirani(2005)*提出了*Fused Lasso*，用于处理有序变量数据结构，尤其擅长解决基因组数据和信号数据，其具体形式如下

$$\hat{\beta} = argmin \frac{1}{2} \|y - X\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \sum_{i=2} |\beta_i - \beta_{i-1}| . \tag{3-10}$$

第一个惩罚项是常见的l_1范数，其目的与*Lasso*模型一致，压缩系数β 趋向于*0*。第二个惩罚项是针对有序的数据情况下，促使相邻系数趋于一致，而且导致一些系数一致。将

上述公式转化为稀疏框架下形式，当

$$T_1 = \begin{bmatrix} I \\ \lambda F_1 \end{bmatrix} \ with\, F_1 = \begin{bmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \cdots & \cdots & \\ & & & 1 & -1 \end{bmatrix}. \tag{3-11}$$

则*Fused Lasso*即转化为稀疏框架下的形式。其内容与*Fused Lasso*完全一致。

Example 3.1.3 (Clustered Lasso) *She(2010)*提出了*Clustered Lasso,*其主要目的是即保持有*Lasso*模型的变量选择，又能够利用聚类的思想将变量进行分组，其具体形式如下

$$\hat{\beta} = argmin \frac{1}{2} \|y - X\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \sum_{i<j} \left|\beta_i - \beta_j\right|. \tag{3-12}$$

将上述公式转化为稀疏框架下行驶，当

$$T_2 = \begin{bmatrix} I \\ \lambda F_2 \end{bmatrix}. \tag{3-13}$$

其中F_2为成对差分矩阵。形式如下

$$F_2(i,j) \begin{cases} 1, & if\ j = \alpha_i, \\ -1, & if\ j = \beta_i, \\ 0, & other. \end{cases} \tag{3-14}$$

其中$i = 1, \cdots, d(d-1)$,d为真实β的个数。$\{(\alpha_i, \beta_i)\}$ 列举了所有可能的成对组合。当$d = 4$时，$F_2 = \begin{bmatrix} 1 & -1 & & & \\ 1 & & -1 & & \\ 1 & & & -1 & \\ & 1 & -1 & & \\ & 1 & & -1 & \\ & & 1 & -1 & \end{bmatrix}$。*Clustered Lasso*虽然与*Fused Lasso* 形式上相似，但他不需要对回归特征进行排序，其试图组织无序的特征并产生结果排序，实际上*Clustered Lasso*是一种监督聚类方法，同时考虑x 和y 来为特征寻找合适的分组。

总而言之，可定制的T表示在回归分析中提出的稀疏性要求，用于表示真实β 之间的特定关系。其表现形式可以是多种多样，而本文的主要研究方向就是基于这种稀疏框架的形式，融合先验信息形成一种新的特定模型。

3.2 基于先验稀疏框架的Lasso 回归的定义

上一节中本文详细介绍了基于稀疏框架的Lasso回归模型，了解了其比较常见的几种特定形式，但是这几个模型均没有利用到先验信息。因此，本文提出了一种新的Lasso改进方法，即将先验信息融入Lasso模型，本文将其称为基于先验稀疏框架的Lasso回归模型。

如果把在回归分析中特征的自有属性看做一个整体，那么先验信息就是该整体中的一部分。我们在建立Lasso回归模型时，有时会提前知道它内在的一些属性及其作用。并且先验信息本身就是存在的，所以与搜集数据无关，反之正确搜集数据应该受到先验信息限制，并且满足先验信息。但是先验信息不能用于描述研究对象，所以需要将特征本身具有的先验信息应用到模型建立中，并通过模型的结果来表现出先验信息。

现在考虑稀疏框架下的Lasso回归，为了更好的表达，将公式(3-9)改为拉格朗日形式表示

$$\hat{\beta}^{TL} = argmin\frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|T\beta\|_1 . \tag{3-15}$$

于此同时为了便于描述融合先验信息后的新模型，本文在此举个例子使读者更好的理解。假设我们获得参数β具有如下先验信息：3个不同的$(\beta_1,\beta_2,\beta_3)$ 是一致的，$(\beta_4 = \beta_5)$, 并且$\beta_6 = 0$。为了满足这种先验，我们可以构造T包含

$$\begin{bmatrix} 1 & 1 & -2 & & & \\ & & & 1 & -1 & \\ & & & & & 1 \end{bmatrix}. \tag{3-16}$$

以便在回归拟合中更好的获得真实的β。但是就算T包含上诉行向量，仍然可以构造不同的T 矩阵。但主要有以下两种结构。

稀疏矩阵T为方阵：假设我们的稀疏框架T为方阵时，其具体形式如下

$$T=\begin{bmatrix} 1 & 1 & -2 & & & \\ & 1 & -1 & & & \\ & & 1 & & & \\ & & & 1 & -1 & \\ & & & & 1 & \\ & & & & & 1 \end{bmatrix}. \tag{3-17}$$

稀疏矩阵T为“瘦”矩阵：假设我们的稀疏框架T不为方阵时，即构造的形式类似Fused Lasso。

$$T=\begin{bmatrix} 1 & & & & & \\ & 1 & & & & \\ & & 1 & & & \\ & & & 1 & & \\ & & & & 1 & \\ & & & & & 1 \\ 1 & 1 & -2 & & & \\ & 1 & -1 & & & \\ & & & 1 & -1 & \end{bmatrix}. \tag{3-18}$$

因此，在后文的求解和实验中，本文会就两种不同形式的T矩阵进行分析比较。并且就先验信息而言，信息高低也会导致结果的优劣。

3.3 基于先验稀疏框架的Lasso 回归的求解

首先，我们引出上一章的Lasso回归模型，给定(X,y,λ),则

$$\hat{\beta}^L = argmin\frac{1}{2}\|y-X\beta\|_2^2+\lambda\|\beta\|_1. \tag{3-19}$$

在满足KKT的条件，$\hat{\beta}^L$是最优解当且仅当$\hat{\beta}^L$满足

$$X^T(X\hat{\beta}^L-y)+\lambda s(\hat{\beta}^L=0, or\ \lambda s(\hat{\beta}^L)=X^Ty-\Sigma\hat{\beta}^L. \tag{3-20}$$

令$\xi=\hat{\beta}^L+\lambda s(\hat{\beta}^L)$。则当$\hat{\beta}^L>0$时，$\xi=\hat{\beta}^L+\lambda;\hat{\beta}^L=0$，$\xi=0$；$\hat{\beta}^L<0,\xi=\hat{\beta}^L-\lambda$。引用前文的软阈值符号，则

$$\hat{\beta}^L=\eta_s(\xi,\lambda). \tag{3-21}$$

将公式(3-10)重写为

$$\hat{\beta}^L+\lambda s(\hat{\beta}^L)=X^Ty+(I-\Sigma)\hat{\beta}^L. \tag{3-22}$$

这就激发了迭代的设计来解决公式(3-9)

$$\xi^{(j+1)}=X^Ty+(I-\Sigma)(\hat{\beta}^L)^{(j)},\ (\hat{\beta}^L)^{(j+1)}=\eta_s(\xi^{(j+1)},\lambda). \tag{3-23}$$

其中j表示迭代次数，初始值的选定可以用随机变量，也可以选用最小二乘估计作为初始值进行迭代。当$\xi^{(j+1)}\approx\xi^{(j)}$时，我们默认迭代结束。该迭代过程收敛到一个最优点。此迭代方式已经以不同的形式提出，并被运用于解决大数据问题。

假定稀疏矩阵T列满秩，因此T存在两种形式，一种为常见的方阵，另一种形如$T=\begin{bmatrix}I\\F\end{bmatrix}$的"瘦"矩阵。借鉴Lasso的求解方法，在满足KKT 的条件，$\hat{\beta}^{TL}$是最优解当且仅当$\hat{\beta}^{TL}$满足

$$X^T(X\hat{\beta}^{TL}-y)+\lambda T^Ts(T\hat{\beta}^{TL})=0. \tag{3-24}$$

或者是

$$\lambda T^Ts(T\hat{\beta}^{TL})=X^Ty-\Sigma\hat{\beta}^{TL}. \tag{3-25}$$

其中$\Sigma=X^TX$,之后将公式两边同时加上$T^TT\hat{\beta}^{TL}$, 则公式(3-25)改写为

$$T^T(T\hat{\beta}^{TL}+\lambda s(T\hat{\beta}^{TL}))=X^Ty+(T^TT-\Sigma)\hat{\beta^{TL}}. \tag{3-26}$$

令$\xi=T\hat{\beta}^{TL}+\lambda s(T\hat{\beta}^{TL})$, 将公式(3-26) 重写为

$$T^T\xi=X^Ty+(T^TT-\Sigma)\hat{\beta}^{TL}. \tag{3-27}$$

若想沿用迭代的思想需要将T^T移至公式右边，那就要求T^T存在逆矩阵。很明显,当稀疏矩阵T为"瘦"矩阵时，矩阵T^T没有左逆。就像fused Lasso 一样，因此本文需要对稀疏矩阵T的两种不同情况进行分类讨论。

3.3.1 稀疏框架T为方阵

首先，我们重置参数，令$\gamma = T\beta$,当稀疏框架T为方阵时，T^{-1} 存在。接下来将拉格朗日的一般形式Lasso(3-9) 重写为

$$\hat{\gamma} = argmin\frac{1}{2}\left\|y - XT^{-1}\gamma\right\|_2^2 + \lambda\left\|\gamma\right\|_1. \tag{3-28}$$

满足KKT条件，$\hat{\gamma}$是最优解当且仅当$\hat{\gamma}$满足

$$(XT^{-1})^T(XT^{-1}\hat{\gamma} - y) + \lambda s(\hat{\gamma}) = 0. \tag{3-29}$$

为了便于阅读，令$\kappa = XT^{-1}$, 带入公式(3-29)得到与Lasso近似的求解方式

$$\kappa^T(\kappa\hat{\gamma} - y) + \lambda s(\hat{\gamma}) = 0. \tag{3-30}$$

或者是

$$\lambda s(\hat{\gamma}) = \kappa^T y - \kappa^T\kappa\hat{\gamma}. \tag{3-31}$$

令$\xi = \hat{\gamma} + \lambda s(\hat{\gamma})$,则

$$\hat{\gamma} = \eta_s(\xi, \lambda). \tag{3-32}$$

重写公式(3-31)

$$\hat{\gamma} + \lambda s(\hat{\gamma}) = \kappa^T y + (I - \kappa^T\kappa)\hat{\gamma}. \tag{3-33}$$

同理沿着迭代算法得出

$$\xi^{(j+1)} = \kappa^T y + (I - \kappa^T\kappa)(\hat{\gamma})^{(j)},\ (\hat{\gamma})^{(j+1)} = \eta_s(\xi^{(j+1)}, \lambda). \tag{3-34}$$

并且最终得出

$$(\hat{\beta}^{TL})^{j+1} = T^{-1}(\hat{\gamma})^{j+1}. \tag{3-35}$$

在稀疏框架T为方阵时，求解算法与标准Lasso的求解算法没有太大区别。其算法步骤如下：

1. 初始化参数，设置初始$\hat{\gamma}^0, \lambda$。

2. 开始迭代:

- 通过公式(3-35)更新$\hat{\gamma}^{(j)}$ 和$(\hat{\beta}^{TL})^{(j)}$。
- 如果$\left\|(\hat{\beta}^{TL})^{(j)}-(\hat{\beta}^{TL})^{(j-1)}\right\|$ 足够小，停止迭代。
- 否则，令$j \to j+1$:继续下一次迭代。

3.3.2 稀疏框架T为"瘦"矩阵

刚才介绍完稀疏框架T为方阵的求解方法，接下来看看当稀疏框架T为"瘦"矩阵的解法。同样，我们重置参数，引入H满足$HT=I$，假定T的SVD分解为$T=UDV^T$，我们可以得出$H=VD^{-1}U^T$。因此公式(3-28)等价以下的Lasso模型

$$\hat{\gamma}=argmin\frac{1}{2}\|y-XH\gamma\|_2^2+\lambda\|\gamma\|_1,\ s.t.\ TH\gamma=\gamma. \tag{3-36}$$

满足KKT条件，$\hat{\gamma}$是最优解当且仅当$\hat{\gamma}$满足

$$(XH)^T(XH\hat{\gamma}-y)+\lambda s(\hat{\gamma})=0. \tag{3-37}$$

展开公式(3-37)

$$H^T\Sigma H\hat{\gamma}-H^TX^Ty+\lambda s(\hat{\gamma})=0. \tag{3-38}$$

或者是

$$\hat{\gamma}+\lambda s(\hat{\gamma})=H^TX^Ty+(I-H^T\Sigma H)\hat{\gamma}. \tag{3-39}$$

令$\xi=\hat{\gamma}+\lambda s(\hat{\gamma})$,则

$$\hat{\gamma}=\eta_s(\xi,\lambda). \tag{3-40}$$

沿着迭代展开

$$\xi^{(j)}=H^TX^Ty+(I-H^TX^Ty)(\hat{\gamma})^{(j-1)},\ (\hat{\gamma})^{(j)}=\eta_s(\xi^{(j)},\lambda). \tag{3-41}$$

且

$$\begin{cases}\hat{\gamma}^{(j+1)}=TH\hat{\gamma}^{(j)},\\(\hat{\beta}^{TL})^{(j+1)}=H\hat{\gamma}^{(j+1)}.\end{cases} \tag{3-42}$$

在某些温和条件下，该算法收敛。其算法步骤如下：

1. 初始化参数，设置初始$\hat{\gamma}^0,\lambda$ 等。

2. 开始迭代:

- 通过公式(3-42)更新$\hat{\gamma}^{(j)}$ 和$(\hat{\beta}^{TL})^{(j)}$。
- 如果$\left\|(\hat{\beta}^{TL})^{(j)}-(\hat{\beta}^{TL})^{(j-1)}\right\|$ 足够小，停止迭代。
- 否则，令$j \to j+1$:继续下一次迭代。

在第四节中，本文会将先验信息放入稀疏框架中，比较两种算法和原始Lasso模型之间的优劣。

3.4 基于先验稀疏框架的Lasso 回归的性质

3.4.1 估计量的相合性

在第二节中本文介绍了Lasso回归估计量的一致性，那么对于稀疏框架下的Lasso 回归，是否具有相同的性质，下面对其进行证明。

首先我们做出如下假设：(1)$\Sigma^n = \frac{1}{n}X_n^T X_n \to \Sigma$，且$\Sigma^n$非奇异或者列满秩。(2)$\frac{1}{n}\underset{1\le i\le n}{max}\, x_i^T x_i \to 0$。(3)$\lambda = o(n)$。

$\hat{\beta}^{TL}$是稀疏框架下的Lasso回归的最优解当且仅当$\hat{\beta}^{TL}$ 满足

$$X^T(X\hat{\beta}^{TL}-y)+\lambda T^T s(T\hat{\beta}^{TL})=0. \tag{3-43}$$

等价于，

$$\hat{\beta}^{TL}=\frac{1}{n}\Sigma^{-1}(X^T y-\lambda T^T s(T\hat{\beta}^{TL})). \tag{3-44}$$

或者是，

$$\hat{\beta}^{TL}=\beta+\frac{1}{n}\Sigma^{-1}X^T\varepsilon-\frac{\lambda}{n}\Sigma^{-1}T^T s(\hat{\beta}^{TL}). \tag{3-45}$$

由Lasso估计量的相合性和假设条件可知：$\frac{1}{n}\Sigma^{-1}X^T\varepsilon \sim N(0,\frac{\Sigma^{-1}}{n}\sigma^2)=O(\frac{1}{\sqrt{n}})=o(1)$，$\frac{\lambda}{n}\Sigma^{-1}T^T s(\hat{\beta}^{TL})=\frac{\lambda}{n}O_p(1)=o_p(1)$。很明显一致性是一个很弱的要求，对于$\Sigma$和$T$没有任何限制，可以很轻易的通过选择合适的$\lambda$ 来实现。因此得出结论，只要$\lambda=o(n)$，则$\hat{\beta}\xrightarrow{p}\beta$, 并且$T\hat{\beta}\xrightarrow{p}T\beta$。

3.4.2 变量选择一致性

在前一节中本文详细的介绍了变量选择一致性和符号一致性，并且后者是明显高于前者的，为了能够有效的选择出在真实β具有相同位置的非零项，因此只要说明稀疏框架下的Lasso 回归模型具有符号一致性就能表示其具有稀疏性，即需要证明$P(sgn(T\hat{\beta}_i)=sgn(T\beta_i))\to 1$。

其实很明显，估计量的相合性就说明了变量选择的一致性。简单来说，对于$\lambda = o(n)$ 而言，我们从上一节能得出$\hat{\beta} \xrightarrow{p} \beta$, 并且$T\hat{\beta} \xrightarrow{p} T\beta$。那么对于前文引入的符号选择性$P(sgn(\hat{\beta}_i) = sgn(\beta_i)) \to 1$，则可以很轻易的说明$P(sgn(T\hat{\beta}_i) = sgn(T\beta_i)) \to 1$。

因此就变量选择一致性而言，只需要控制$\lambda = o(n)$这个控制条件，稀疏框架下的Lasso 回归模型同样具有变量选择一致性和相合性。

3.5 本章小结

针对Lasso回归模型的改进和扩展的方法有很多，本文引入了一种更为宽泛的Lasso回归模型，即基于稀疏框架的Lasso回归，并在此之上融入了先验信息的思想。紧接着给出了新模型的具体形式，并且就两种不同形式的T 给出了相应的求解算法，最后说明了新模型所具有的良好性质，为后文的数值实验做出理论支撑。

第4章　随机模拟与实例展示

为了说明基于先验稀疏框架的Lasso回归模型的性能是优于普通Lasso回归模型的，在我们将传统Lasso 模型和新模型在模拟和实例数据上进行研究。具体地，在模拟实验中我们计算并比较各个模型的均方误差和预测误差，并且画图描述了各模型计算的参数值，在实例分析中我们计算不同模型的预测误差。

4.1　数据模拟

我们在多个仿真数据集上进行了实验。具体数据由以下方式生成

$$y = X\beta + \sigma\varepsilon\ , \varepsilon \sim N(0,1). \tag{4-1}$$

每一个数据集均包含训练集和测试集。统一设定测试集数量为样本数量的30%，令Σ 为生成X 的相关性矩阵，即X 的每一行独立于$N(0,\Sigma)$。我们使用$(\{a_1\}^{n_1},\cdots,\{a_k\}^{n_k})$来表示参数$\beta$ 的个数和具体数值。

并且为了更形象的描述先验信息的多寡，我们定义$knows$ =已知参数关系/总参数个数。举个例子，$\beta = (\{-2\}^{10},\{2\}^{10},\{-4\}^{10},\{4\}^{10},\{0\}^{10})$, 假设我们知道10个$\{-2\}$的比例关系，即在50个参数中，已知10个参数的关系，则记$knows = 20\%$。并且在接下来的实验中，我们均假定已知参数间的关系为1:1形式，即在上面例子中，10个$\{-2\}$ 均相等。具体模型参数如表4.1所示：

表 4.1: 模型参数设置

模型	n	p	σ	$knows$	β
1	100	50	3	0.1	$(\{-2\}^5,\{2\}^5,\{-4\}^5,\{4\}^5,\{0\}^{30})$
2	100	50	6	0.1	$(\{-2\}^5,\{2\}^5,\{-4\}^5,\{4\}^5,\{0\}^{30})$
3	100	50	3	0.3	$(\{-2\}^5,\{2\}^5,\{-4\}^5,\{4\}^5,\{0\}^{30})$
4	100	50	6	0.3	$(\{-2\}^5,\{2\}^5,\{-4\}^5,\{4\}^5,\{0\}^{30})$
5	100	100	3	0.1	$(\{-2\}^{10},\{2\}^{10},\{-4\}^{10},\{4\}^{10},\{0\}^{60})$
6	100	100	6	0.1	$(\{-2\}^{10},\{2\}^{10},\{-4\}^{10},\{4\}^{10},\{0\}^{60})$
7	100	100	3	0.3	$(\{-2\}^{10},\{2\}^{10},\{-4\}^{10},\{4\}^{10},\{0\}^{60})$
8	100	100	6	0.3	$(\{-2\}^{10},\{2\}^{10},\{-4\}^{10},\{4\}^{10},\{0\}^{60})$
9	100	400	3	0.1	$(\{-2\}^{40},\{2\}^{40},\{-4\}^{40},\{4\}^{40},\{0\}^{240})$
10	100	400	6	0.1	$(\{-2\}^{40},\{2\}^{40},\{-4\}^{40},\{4\}^{40},\{0\}^{240})$
11	100	400	3	0.3	$(\{-2\}^{40},\{2\}^{40},\{-4\}^{40},\{4\}^{40},\{0\}^{240})$
12	100	400	6	0.3	$(\{-2\}^{40},\{2\}^{40},\{-4\}^{40},\{4\}^{40},\{0\}^{240})$

表4.1展示了Lasso模型所有的参数设置。在自变量X是否独立的情况下，设置参数n均为100, 参数p 分别为50、100、400，分别代表低维数据、常规数据、高维数据情况下的模拟。在处理上文所示的数据模拟时，本文均通过*Python*语言来实现模型的构造和数据的处理(包括X的标准化和y 的中心化)。在上一节中我们介绍了T矩阵具有两种不同的形式，因此在对上述数据进行模拟时，分别记*PI*1_*Lasso*和*PI*2_*Lasso* 为基于先验稀疏框架下的Lasso模型，*PI*1_*Lasso*的矩阵形式为公式(3-17)，即稀疏矩阵T 为方阵。*PI*2_*Lasso* 的矩阵形式为公式(3-18)，即稀疏矩阵T 为"瘦"矩阵。

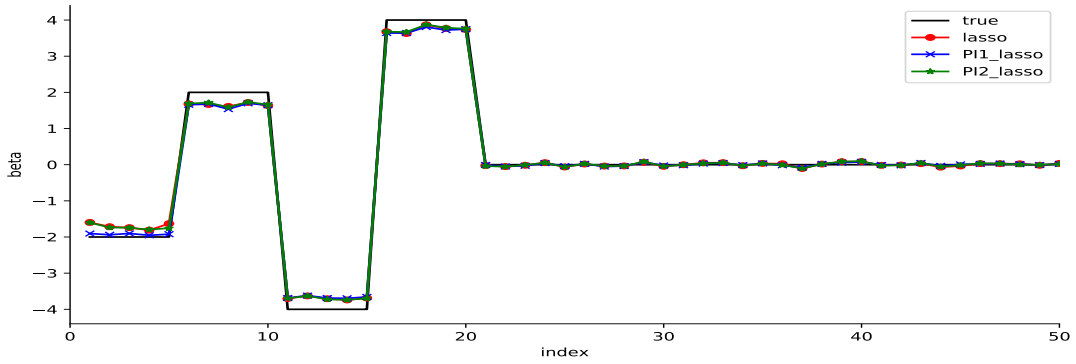
对于上述每个模型均进行了50次模拟，并通过平均的均方误差和平均的预测误差对各方法的性能进行了测试。

表 4.2: 低维数据性能测试结果

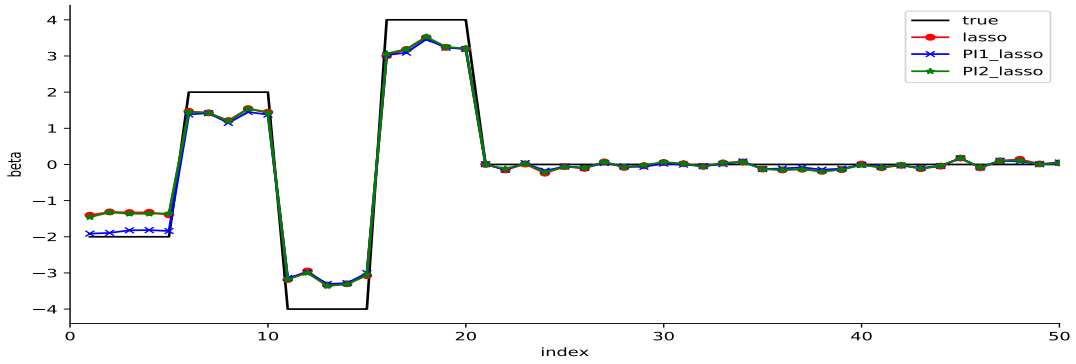
模型	*Lasso*		*PI*1_*Lasso*		*PI*2_*Lasso*	
	MSE	*PE*	*MSE*	*PE*	*MSE*	*PE*
			x_i 相互独立			
1	0.250	21.850	0.194	19.299	0.231	21.097
2	0.899	87.359	0.739	78.207	0.824	83.455
3	0.248	22.745	0.082	13.428	0.189	19.711
4	0.905	82.560	0.368	57.582	0.751	76.344
			x_i 不独立			
1	0.297	18.077	0.201	15.357	0.257	17.132
2	0.844	72.130	0.660	64.832	0.765	68.640
3	0.261	18.055	0.061	12.373	0.196	16.331
4	0.881	70.431	0.235	49.041	0.659	64.639

注：X不独立且服从$\Sigma_{ij} = \rho^{|i-j|}, with\ \rho = 0.5$

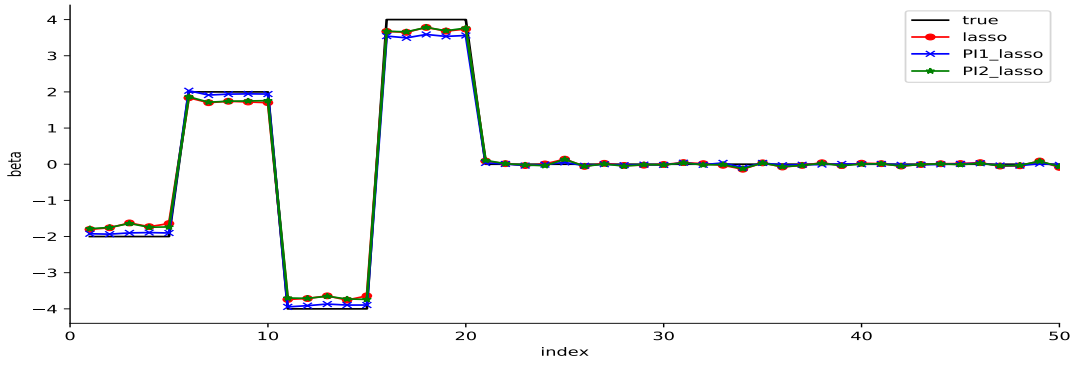
表4.2展示了50次模拟中低维数据各模型的平均均方误差和平均预测误差。从总体来看，基于先验稀疏框架的Lasso模型各性能均显著优于传统的Lasso模型，并且就新模型的两种构造方法而言，*PI*1_*Lasso*模型的性能要优于*PI*2_*Lasso* 模型，这可能与*PI*2_*Lasso*的求解算法有关，引用传统的坐标下降法来求解该模型无法收敛到最优点，这点在She(2010) 的文章中也有进行相应的解释。对于X 是否独立这一条件而言，对于新模型而言影响不大，总体的性能与原Lasso 模型同等降低。接着我们对比参数$\sigma = 3/6$ 的情况，在低维数据中，当$\sigma = 6$，无论是*PI*1_*Lasso*还是*PI*2_*Lasso*模型，性能没有在$\sigma = 3$ 时提升的那么显著。最后对比先验参数$know = 10\%/30\%$ 的情况，当已知先验信息增加时，新模型的性能提升效果非常明显，并且*PI*1_*Lasso*模型的性能提升要显著于*PI*2_*Lasso*模型。



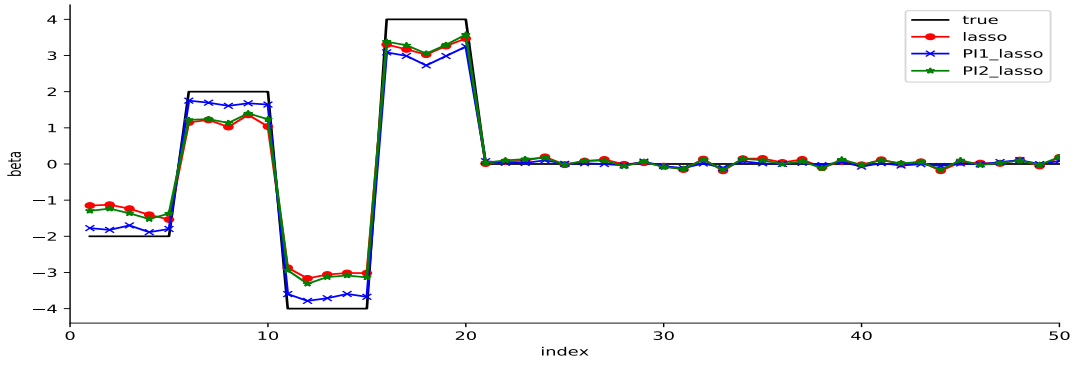
(a) Model1



(b) Model2

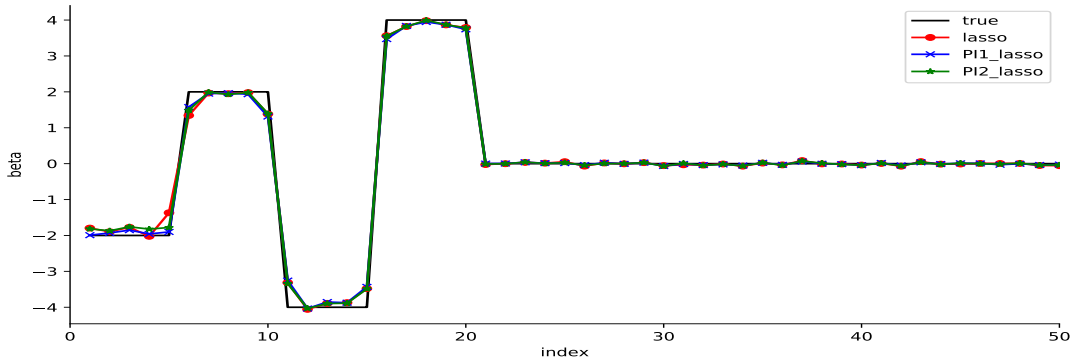


(c) Model3

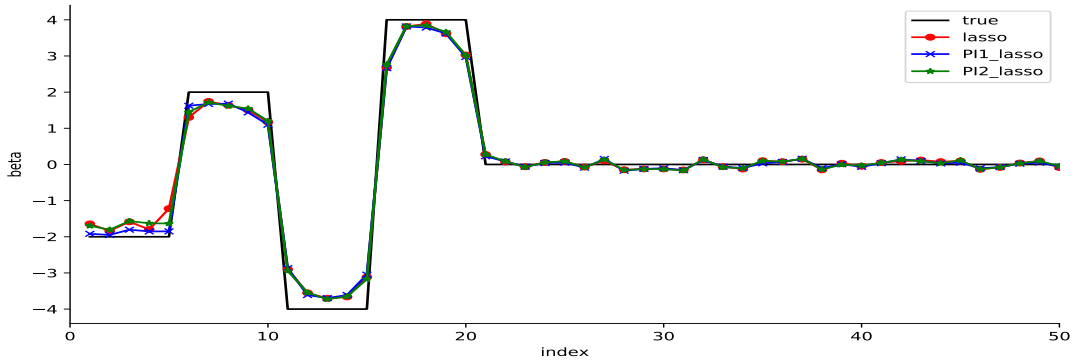


(d) Model4

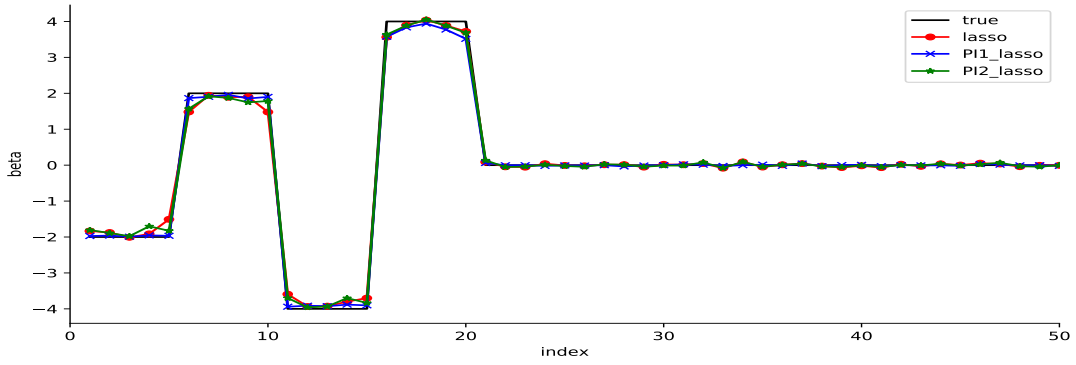
图 4.1: 低维数据，X相互独立的情况下，黑色、红色、蓝色、绿色分别代表真实的β、原$Lasso$模型、$PI1_Lasso$模型、$PI2_Lasso$模型。



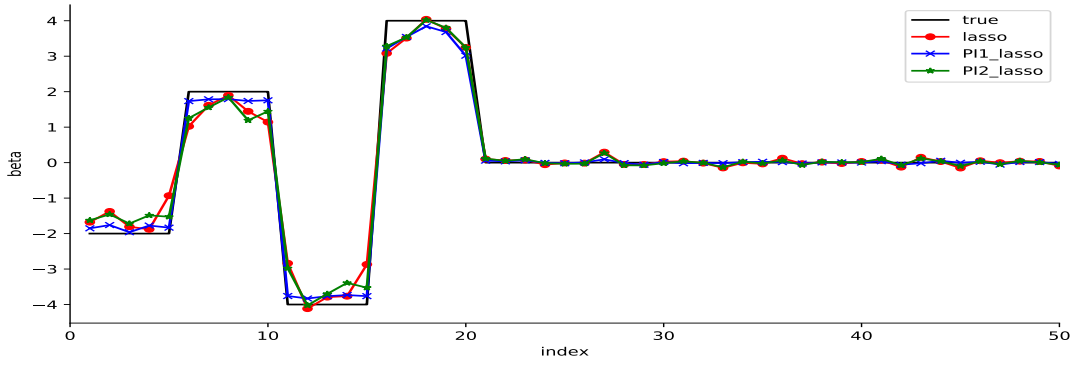
(a) Model1



(b) Model2



(c) Model3



(d) Model4

图 4.2: 低维数据，X不独立的情况下，参数的估计结果展示。

图4.1和图4.2展示了低维数据的参数对比结果。黑色实线代表真实的β，红色$\bullet$ 线代表原Lasso模型的参数估计结果，蓝色$*$线代表$PI1_Lasso$ 模型的参数估计结果，绿色$\star$线代表$PI2_Lasso$模型的参数估计结果。对比发现，在低维数据中，$PI1\_Lasso$ 模型拟合的参数与真实β最为接近，并且提升部分正是我们已知的先验部分。$PI2\_Lasso$ 模型拟合的参数略优于原Lasso 模型。接着我们来看看在常规数据中数据模拟的结果如何，具体模拟结果如下所示：

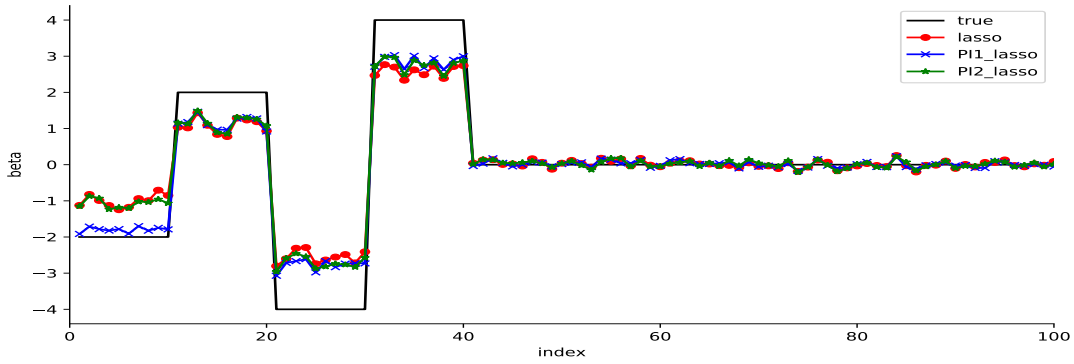
表 4.3: 常规数据性能测试结果

模型	$Lasso$		$PI1\_Lasso$		$PI2\_Lasso$	
	MSE	PE	MSE	PE	MSE	PE
			x_i 相互独立			
5	1.714	189.242	1.05	127.938	1.462	169.264
6	2.38	267.259	1.783	218.882	2.18	251.205
7	1.556	165.9	0.132	23.274	1.084	120.331
8	2.205	268.092	0.482	88.469	1.691	217.199
			x_i 不独立			
5	0.981	89.889	0.522	53.278	0.824	78.053
6	1.686	172.062	1.245	131.391	1.607	162.259
7	0.874	80.192	0.066	15.127	0.550	55.209
8	1.675	169.902	0.255	62.936	1.092	127.909

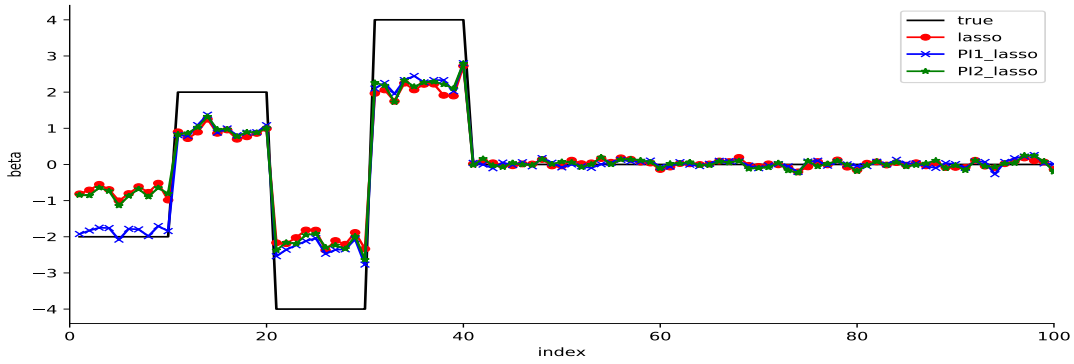
注：X不独立且服从$\Sigma_{ij} = \rho^{|i-j|}, with\ \rho = 0.5$

表4.3展示了50次模拟中常规数据各模型的平均均方误差和平均预测误差，与低维数据类似，从总体来看，基于先验稀疏框架的Lasso模型各性能也显著优于传统的Lasso模型，并且就新模型的两种构造方法而言，$PI1\_Lasso$ 模型的性能要优于$PI2_Lasso$ 模型。不过相较于低维数据而言，$PI1_Lasso$模型的性能提升更加显著，如模型7，,平均均方误差缩小为原Lasso模型的1/10。而在低维数据中，提升效果最好的模型3，平均均方误差也只缩小为原Lasso模型的1/5。类似的，$PI2\_Lasso$ 模型中提升的最明显缩小到原Lasso 模型的2/3, 而在低维数据中，提升效果最好的模型3，平均均方误差也只缩小为原Lasso模型的8/9。

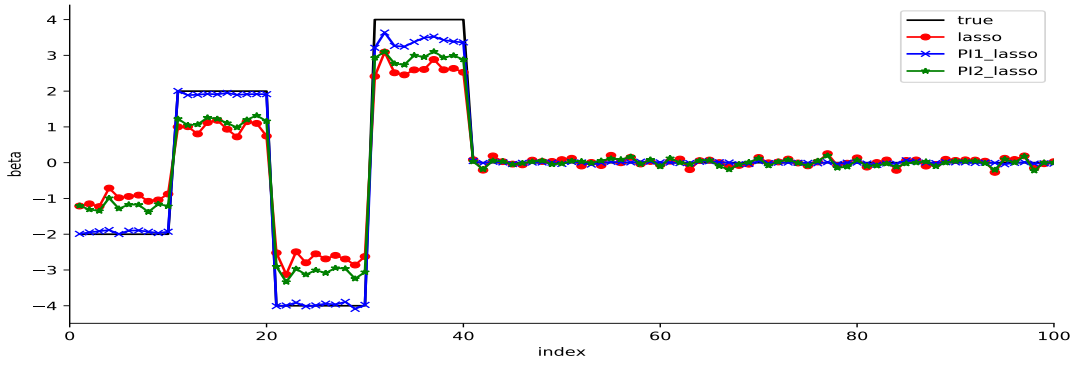
图4.3和图4.4展示了常规数据的参数估计情况，想较于低维数据，图片很好的展示了新模型参数估计的优势。在X 相互独立的情况在，在Model7 中，毋庸置疑，$PI1\_Lasso$ 模型又展示了其良好的参数估计，并且最让人惊喜的，就后一种方法构造的$PI2_Lasso$ 模型，绿色$\star$线相比于红色$\bullet$ 线，也要更贴近与真实的参数β，类似的情况在Model8 中也有表现。



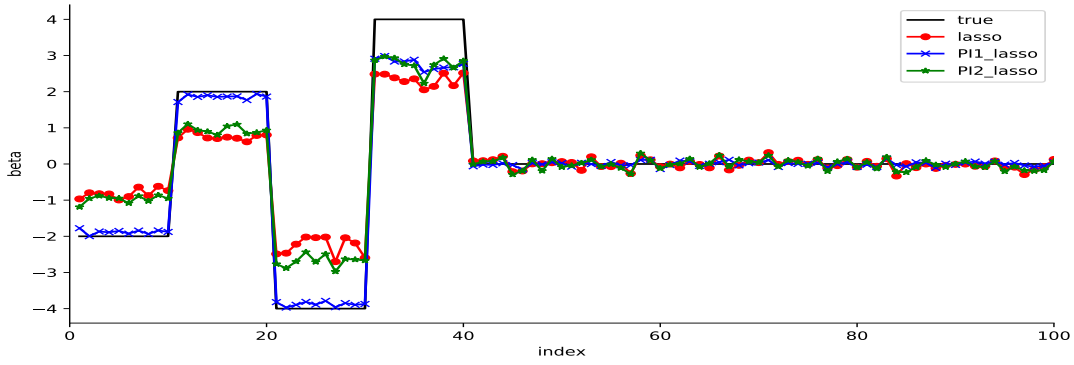
(a) Model5



(b) Model6

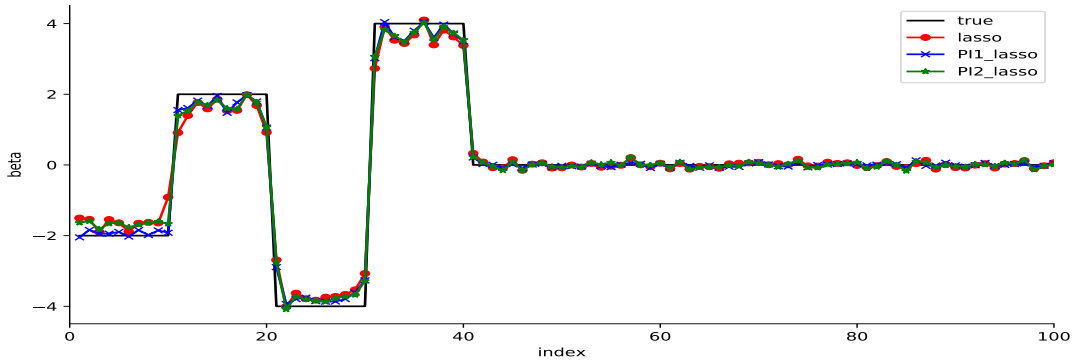


(c) Model7

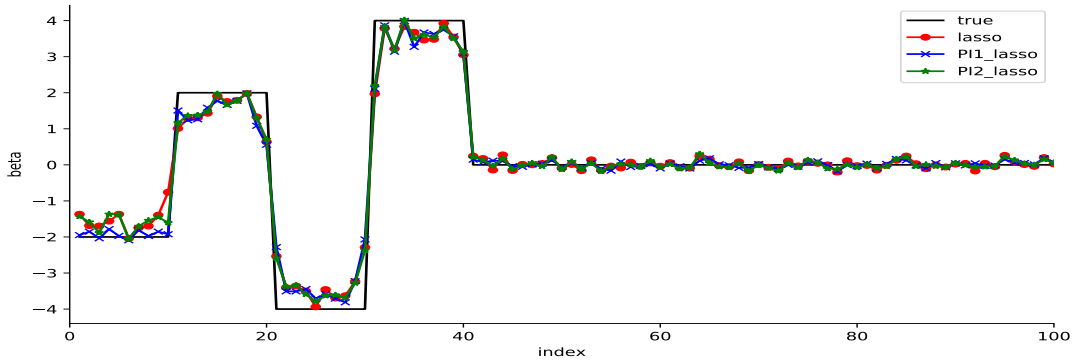


(d) Model8

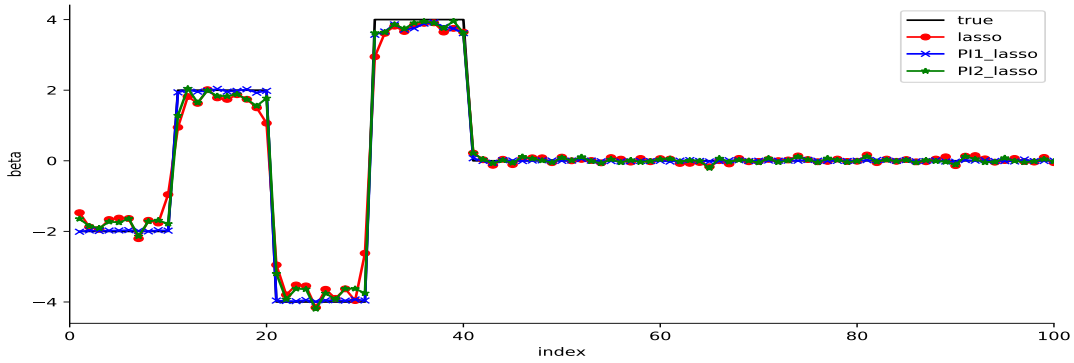
图 4.3: 常规数据，X相互独立的情况下，黑色、红色、蓝色、绿色分别代表真实的β、原$Lasso$模型、$PI1_Lasso$模型、$PI2_Lasso$模型。



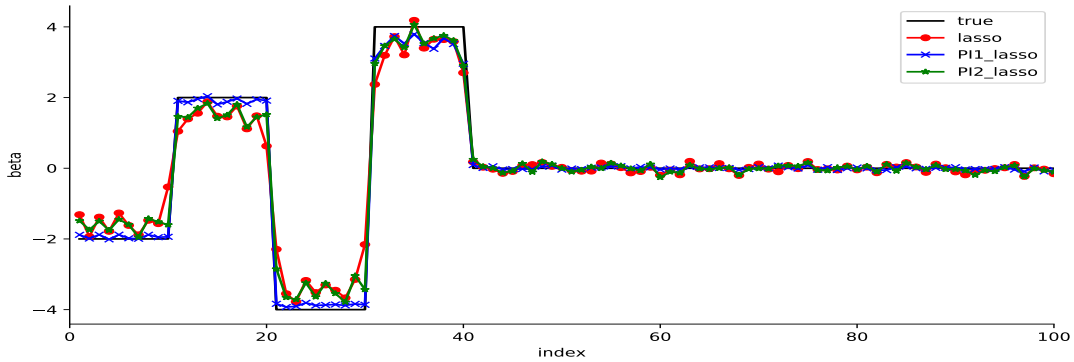
(a) Model5



(b) Model6



(c) Model7



(d) Model8

图 4.4: 常规数据，X不独立的情况下，参数的估计结果展示。

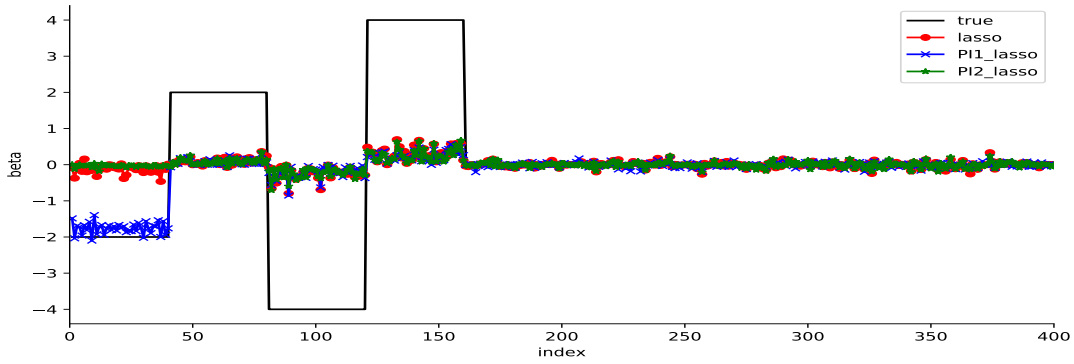
最后，表4.4展示了50次模拟中高维数据各模型的平均均方误差和平均预测误差。无论是在参数$\sigma = 3/6$的情况，亦或者是X是否独立的情况，*PI*1_*Lasso* 模型依然展示了其良好的性能，并且随着先验信息了解的越多，*PI*1_*Lasso*性能的提升就越明显。这一点与前两种状况并没有太大区别。很可惜的是，*PI*2_*Lasso*模型在高维数据前提下，平均均方误差和平均预测误差与原Lasso模型并没有较大差别，甚至在有些模型中，还展示出略逊与原Lasso 模型的性能。不过在先验信息了解增加的情况下，其平均均方误差和平均预测误差依然还是优于原Lasso模型。

表 4.4: 高维数据性能测试结果

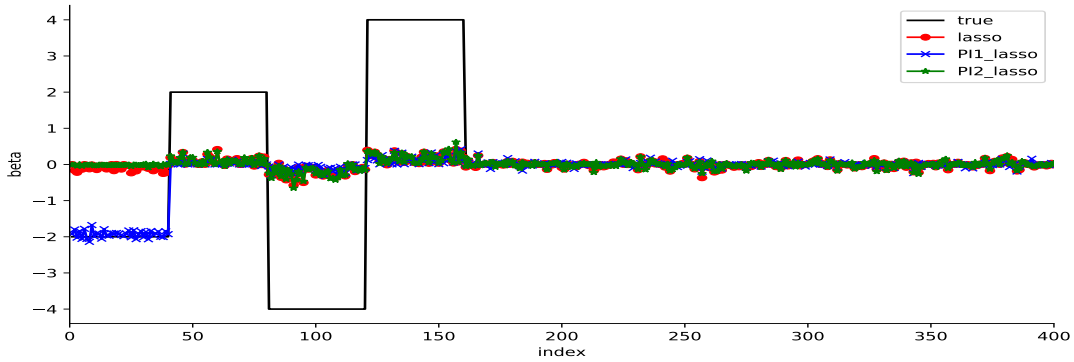
模型	*Lasso*		*PI*1_*Lasso*		*PI*2_*Lasso*	
	MSE	*PE*	*MSE*	*PE*	*MSE*	*PE*
			x_i 相互独立			
9	3.989	1706.42	3.701	1676.48	4.005	1681.42
10	4.076	1951.67	3.785	1844.57	4.062	1907.11
11	3.995	1760.58	1.647	827.79	4.071	1686.69
12	4.015	1746.83	1.660	826.386	4.034	1694.45
			x_i 不独立			
9	4.711	3235.7	4.206	2925.77	4.753	3232.29
10	4.565	3356.81	4.177	2921.72	4.494	3281.5
11	4.575	3444.8	1.239	733.656	4.755	3189.480
12	4.755	3338.38	1.218	691.715	4.725	3187.44

注：X不独立且服从$\Sigma_{ij} = \rho^{|i-j|}, with\ \rho = 0.5$

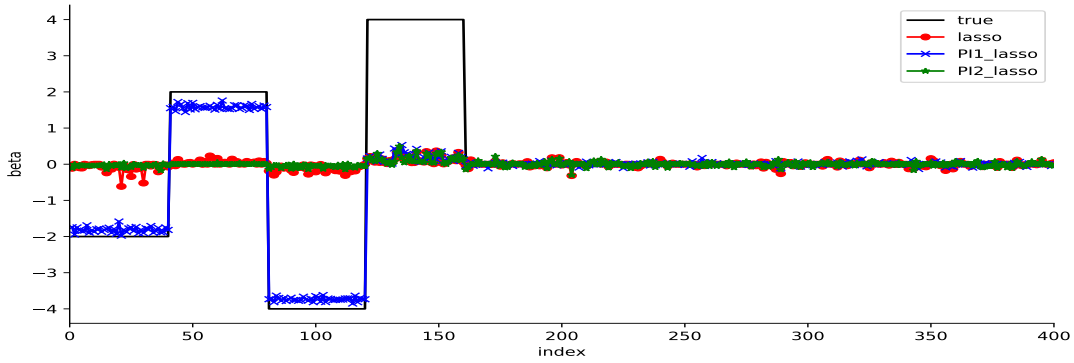
与前两种数据状况类似，图4.5和图4.6展示了高维数据的参数估计情况。*PI*1_*Lasso* 模型依然表现了其良好的参数估计状况，无论是先验信息已知的多寡，在图片中均很好的表现了出来。并且与前面分析的一致，*PI*2_*Lasso* 模型与原Lasso模型虽然在表4.4中可以直观的看出区别，但在图片中二者几乎一致。



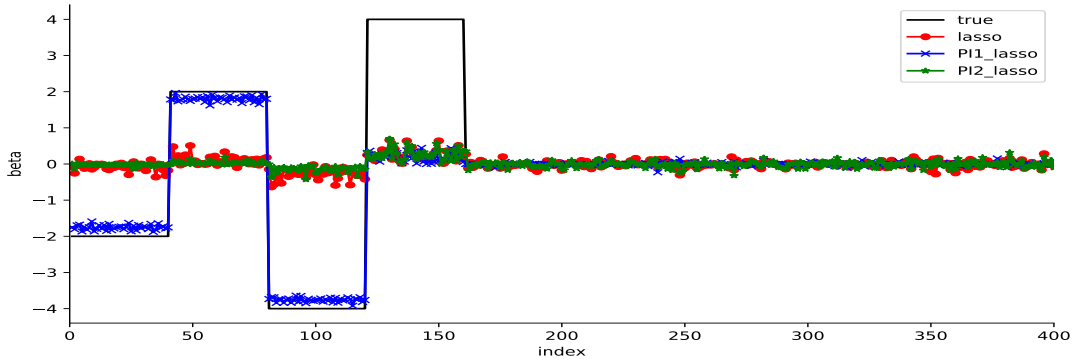
(a) Model9



(b) Model10

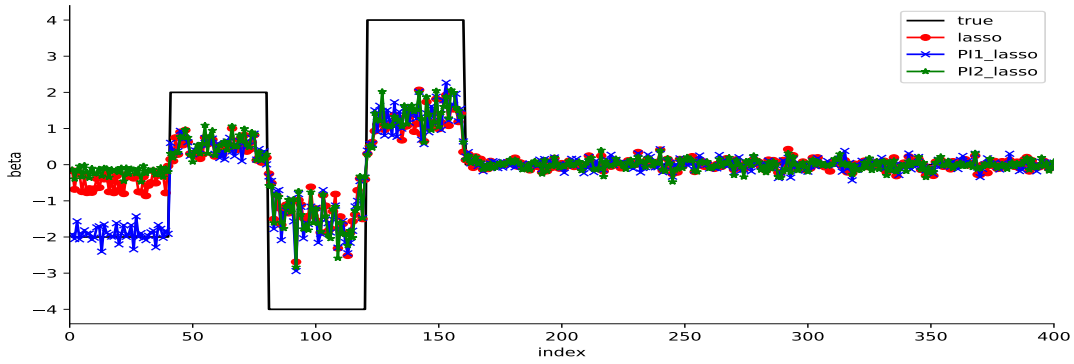


(c) Model11

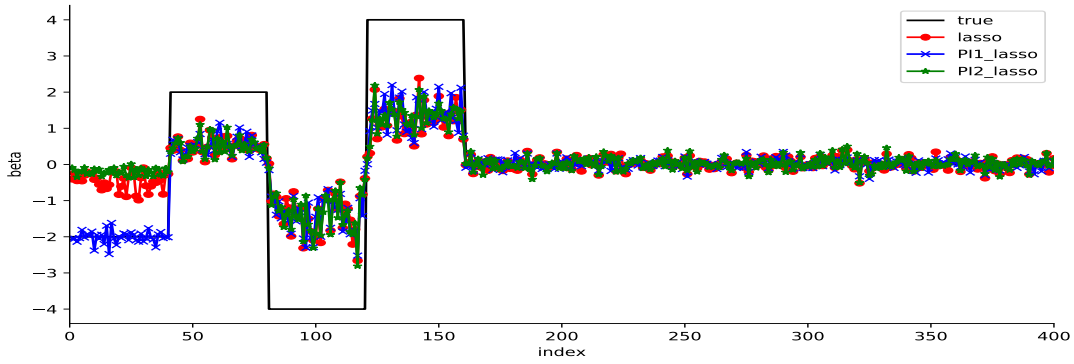


(d) Model12

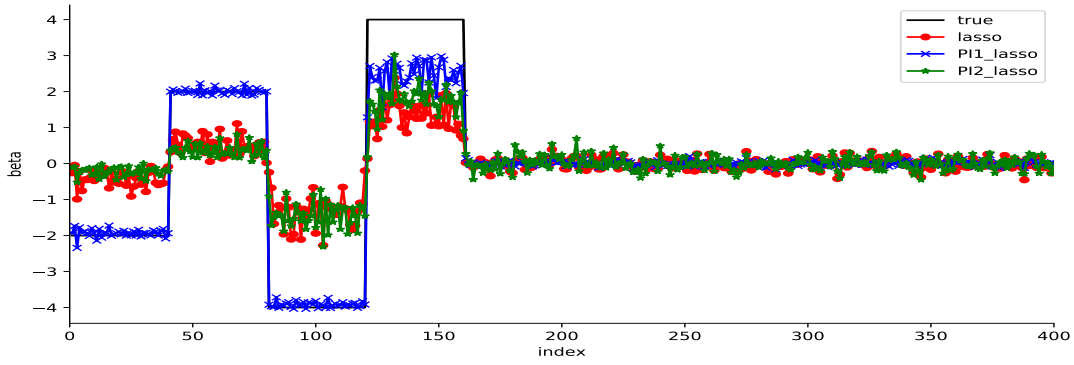
图 4.5: 高维数据，X相互独立的情况下，黑色、红色、蓝色、绿色分别代表真实的β、原*Lasso*模型、*PI*1_*Lasso*模型、*PI*2_*Lasso*模型。



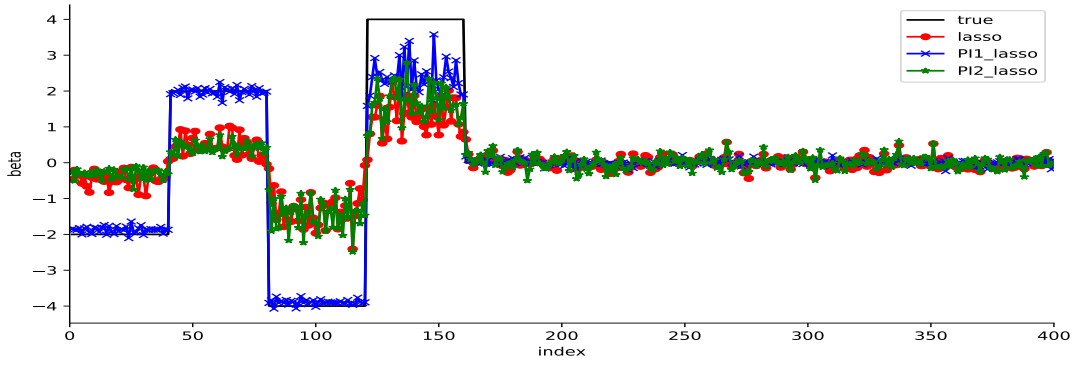
(a) Model9



(b) Model10



(c) Model11



(d) Model12

图 4.6: 高维数据，X不独立的情况下，参数的估计结果展示。

4.2 实证分析

4.2.1 样本数据概况

为了更好的说明本文提出的基于先验稀疏框架的Lasso模型性能是优于原Lasso 模型，我们将在本节中采用真实数据集对其进行实验验证。从大数据竞赛平台*Kaggle* 中获取了NBA 球员数据(https://www.kaggle.com/drgilermo/nba-players-stats)。

该数据集包含67个NBA赛季的个人统计数据，数据信息如表4.5所示：

表 4.5: 数据信息说明

文件	样本量	特征量	说明
player_data.csv	4551	8	记录球员个人信息
Players.csv	3923	7	记录球员个人信息
Seasons_States.csv	24.7k	53	记录球员效力NBA期间的各项数据

其中*player_data.csv*文件记录了各球员信息，包括：姓名、效力开始时间、效力截止时间、位置、大学名称、身高、体重、出生日期8个列特征。

*Players.csv*文件记录各球员类似的信息，包括：姓名、出生省份、出生地区、身高、体重、大学、出生日期7 个列特征。

最重要的*Seasons_States.csv*数据记录了各球员在效力NBA期间的各项数据，包括上场时间、助攻、命中率、年龄、得分、效力队伍等53个列特征。

由于未知原因，该数据集在1950-1973 年间并没有统计进攻篮板率、防守篮板率、总篮板率、出场率、抢断率等13个特征变量。但在1974-1980年间这些特征均有相应的记录。而本实验的目标是为了分析在两分球时代(1980年以前)，各项特征对球员贡献率的影响。为了充分利用完整数据集，本文将利用1950-1973 年的数据分别做Lasso 回归拟合和最小二乘拟合，将观察得出的特征间的比例关系作为先验信息。在此之后利用1973-1980年的数据，分别基于先验稀疏框架的Lasso 模型和原Lasso模型做相应的回归拟合，对比不同模型不同先验情况下的性能差距。

4.2.2 数据处理及实验结果

在进行数据分析和建模的过程中，数据处理是必不可少的步骤，因为往往原始数据并不能完全满足构造模型的要求。同样的，对于本文使用的NBA数据，即使其原作者已经进行了部分适当的处理，但为了适用本文的目标要求，我们仍需要将数据进行相应的处理。在此重申，本文在处理上文所示的数据集时，均通过*Python*来实现模型的

构造和数据的处理。数据处理具体步骤如下：

1. 将三个文件按照球员姓名为关键词合并，选出1950-1980年间的球员信息。
2. 将合并后的文件按照球员姓名分组，以贡献率高低，选出每个球员效力NBA 期间的最佳贡献率。
3. 剔除缺失值较多的变量，包括出生日期、省份和地区。创造新变量“效力时长”：效力时长= 效力截止时间-效力开始时间。
4. 对于少数具有缺失值的变量，以该特征均值进行填充。对于定类的特征变量，本文处理的方式是将其转化为哑变量。最后，我们会将所得到的数据矩阵进行标准化处理。

通过上述数据处理后，我们得到本文需要的样本数据集为1146×291。其中1950-1973年的数据集为691×278，将其记为先验数据集；1974-1980年的数据集为455×291，将其记为实证数据集。并且上述数据集中均有255 个为哑变量。

由于在模拟实验中$PI1_Lasso$形式表现的性能要普遍优于$PI2_Lasso$ 形式，因此在实证分析中，本文只选用$PI1_Lasso$形式的新模型进行实证。并且前文中介绍的先验信息是由最小二乘估计和Lasso估计两种估计方法得到的参数比例，因此我们将两种得出先验的形式分别记为PIO_Lasso 和PIL_Lasso，前者为利用最小二乘得出的先验，后者为利用Lasso 得出的先验。

我们取出先验数据集，将贡献率作为响应变量矩阵Y，去除年份和球员姓名后的数据集作为解释变量矩阵X。利用最小二乘估计进行50次数据拟合。观察拟合参数间的比例关系。

表 4.6: 最小二乘参数估计结果展示

特征	mean	std	median
犯规次数	-0.003	0.001	0.003
上场时间	0.002	$2e^{-4}$	0.002
身高	0.004	0.008	0.004
体重	-0.006	0.005	-0.006
效力时长	0.057	$3e^{-9}$	0.055
年龄	-0.011	0.001	-0.012
罚球命中次数	-0.011	0.002	0.011
罚球命中率	0.001	$6e^{-4}$	$8e^{-4}$

表4.6展示了最小二乘估计在50次模拟中参数估计量的描述性统计，其中mean、std、median 分别代表50次模拟中参数估计值的均值、标准差和中位数。由于考虑参数

太小在利用新模型进行拟合时会压缩至0，本文选取均值>0.001, 且标准差较小的参数估计量。其中包括：犯规次数、上场时间、身高、体重、效力时长、年龄、罚球命中次数、罚球命中率8个特征变量。从上述三个指标看，每个特征均具有良好的代表性。同时为了更简单的表示T矩阵，分别将上述8 个特征表示为$(\beta_1,\cdots,\beta_8)$，并采取两两捆绑的形式，即$\beta_1+\frac{3}{2}\beta_2=0,\beta_3+\frac{2}{3}\beta_4=0,\beta_5+5\beta_6=0,\beta_7+11\beta_8=0$。因此，最小二乘估计形成$T$ 矩阵包括$\begin{bmatrix}1 & \frac{3}{2} & & & & & & \\ & & 1 & \frac{2}{3} & & & & \\ & & & & 1 & 5 & & \\ & & & & & & 1 & 11\end{bmatrix}$。

了解完最小二乘估计拟合的参数，我们继续观察Lasso估计拟合的参数，同最小二乘估计类似，我们取出先验数据集，将贡献率作为响应变量矩阵Y，去除年份和球员姓名后的数据集作为解释变量矩阵X。利用Lasso估计进行50次数据拟合。观察拟合参数间的比例关系。

表 4.7: Lasso估计结果展示

特征	mean	std	median
投中次数	-0.016	$6e^{-4}$	-0.016
犯规次数	0.014	$6e^{-4}$	0.014
上场时间	0.002	$2e^{-4}$	0.002
得分命中率	0.003	0.001	0.003
助攻	-0.002	$9e^{-4}$	0.002

表4.7展示了Lasso估计在50次模拟中参数估计量的描述性统计，其中mean、std、median 分别代表50 次模拟中参数估计值的均值、标准差和中位数。采取与最小二乘相同的标准选取参数估计量，其中包括：投中次数、犯规次数、上场时间、得分命中率、助攻5个特征变量。与最小二乘估计结果类似，5 个特征变量均具有良好的代表性。同样，将上述5 个特征变量表示为$(\beta_1,\cdots,\beta_5)$，由于Lasso估计得到的特征变量较少，因此本文将前两个变量捆绑，后三个变量捆绑。即$\beta_1+\frac{8}{7}\beta_2=0,3\beta_3+\beta_4+4.5\beta_5=0$。因此，Lasso估计形成的$T$矩阵包括$\begin{bmatrix}1 & \frac{8}{7} & & & \\ & & 3 & 1 & 4.5\end{bmatrix}$。

在处理完先验数据集得出先验信息后，我们取出实证数据集，该数据集相比前者增加了13 个特征变量。本文运用先验数据集估计的参数比例关系作为先验信息，利用上节中提出的$PI1_Lasso$模型，分别将最小二乘和Lasso估计得出的先验参数比例关系，融入$PI1_Lasso$ 模型中，分别取名为PIL_Lasso和PIO_Lasso。

表4.8展示了两种先验信息情况下新模型与原Lasso模型的性能比较。由于真实参数

表 4.8: 基于先验稀疏框架的Lasso模型与原Lasso模型的性能比较

模型	*PE*			*MAE*		
	mean	std	median	mean	std	median
Lasso	0.469	0.075	0.469	0.473	0.043	0.475
PIL_Lasso	0.459	0.068	0.448	0.464	0.037	0.462
PIO_Lasso	0.455	0.068	0.450	0.461	0.038	0.458

未知，为了更好的体现新模型的作用，本文在实证部分增加了一个评价指标，即平均绝对误差(Mean Absolute Error)：$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, i = 1,2,\cdots,n$。其中$y_i$ 为响应变量，$\hat{y}_i$ 为根据模型预测的响应变量。同时为了合理的表示真实数据的性能，我们对实证数据集做了50 次试验，上表中mean、std、median分表表示*MSE* 和*MAE*50次试验结果的均值、标准差和中位数。

从表4.8展示的结果可知，基于先验稀疏框架的Lasso模型从均值和中位数上看总体是优于原Lasso模型的，并且稳定性也略优于原Lasso模型。在利用LASSO 估计得出的先验参数比例时，无论是*PE* 上和*MAE* 上均得到了显著的提升，尤其是在先验信息很少的情况下，我们只了解了5个特征的关系，换算成$knows = 5/278 \approx 1.8\%$ 的情况下，*PE* 提升了2.1%，*MAE*提升了1.9%。在利用Lasso 估计得到的先验参数比例，与最小二乘估计类似，我们将先验信息进行换算，$knows = 8/278 \approx 2.8\%$ 的情况下，*PE* 提升了3.2%，*MAE*提升了2.5%。

4.3 本章小结

通过模拟和实证的结果来看，我们得出了以下结论：

1. 基于先验稀疏框架的Lasso 模型总体性能是优于原始Lasso模型，无论先验稀疏矩阵T是方阵还是"瘦"矩阵，因此利用先验稀疏框架的方法具有良好的可行性和一定的推广性。

2. 对于某种原因导致部分样本的部分特征存在缺失时，可以利用基于先验稀疏框架的Lasso模型进行处理，会得到优于原Lasso模型的结果。因此使用该模型存在一定的局限性。

3. 在数据处理中，如果先验信息获取的比例越高，基于先验稀疏框架的Lasso 模型的性能相比原始Lasso模型就越好。

第5章　总结与展望

5.1　研究结论

在数据量呈现爆炸式增长的今天，涌现出了各式各样的数据模型。与此同时，变量选择问题一直是统计学理论和应用中的一个重要问题，因此如何选择一种好的变量选择方法是很有必要的。好的变量选择不仅可以避免维度灾难的问题，还可以保证预测的精准度及模型的稳定性。其中使用最多的模型就是Lasso 模型，相较于传统的回归模型，其主要的优势在于可以对多余的变量进行压缩并剔除影响程度较低的变量，从而很好的达到变量选择的作用。因此，在处理高维数据情况中，Lasso模型具有很好的适用性，并且其改进方法在统计学研究领域中也逐渐受到人们的关注。

本文在查阅了大量的国内外相关文献的基础上，提出了基于先验稀疏框架下的Lasso 回归模型的研究及其实际运用，论文的主要工作总结如下：

1. Lasso模型作为变量选择领域的重要模型，在保留高精准度的前提下起到了变量选择的作用，可以有效的解决维度灾难的问题，因此在本文的开篇深入浅出的阐述了Lasso 回归模型的原理、求解算法及其性质，证明了其在变量选择的优势和特点。

2. 先验信息作为自由属性的一部分，其本身就是存在的，本文提出了一种将先验信息融入Lasso模型的方法，将其称为基于先验稀疏框架下的Lasso 回归模型。紧接着就融合方式提出了两种不同的形式，并就不同形式分别给出了相应的求解算法，并对新模型的性质进行了证明。

3. 文章的最后就新模型的不同形式分别进行了数据模拟和实证分析，无论是高纬、低维还是常规数据情况下，其结果表明基于先验稀疏框架的Lasso 模型总体表现是优于传统Lasso 模型的。并且就融合形式而言，T矩阵为方阵时效果优于T 矩阵为”瘦”矩阵的情况。

5.2　本文展望

本文在结合文献和研究内容后，提出以下需要研究的方向：

1. Lasso模型特别适用于高维数据情况下，而基因数据很符合这种数据结构，并且就基因学中存在基因通路、重叠基因、等位基因等具有先验信息的说法。因此可以

收集符合本文要求的数据进行更多的实证分析，给基因学提供借鉴意义。

2. 实验数据需要先验信息，并且该先验信息能够为我所用，因此该方法具有一定的局限性，并不是所有数据均能用该方法进行实验。如何有效的打破本文数据的局限性，仍是接下来需要解决的方向。

3. 将先验信息融入模型的思想不仅仅存在于Lasso回归模型，该想法可以融入广泛的数据模型，只要该模型具有正则化的结构，均可以将本文的方法进行融合。

4. 从实证分析来看，如果我们获得具有相似结构的数据，即由于种种原因导致部分特征数据大量缺失，均可以利用本文提出的新模型进行处理，该方法给与了数据清洗一种新思路。

参考文献

[1] 陈希孺. 线性模型中的最小二乘法[M]. 上海: 上海科学技术出版社, 2003.

[2] 哈斯蒂, 蒂布希拉尼, 韦恩. 稀疏统计学习及其应用[M]. 刘波,景鹏志译. 北京: 人民邮电出版社, 2018.

[3] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012.

[4] 李金昌, 苏为华. 统计学(第四版)[M]. 北京: 机械工业出版社, 2014.

[5] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.

[6] Afonso M V, Bioucas-Dias J M, Figueiredo. Fast image recovery using variable splitting and constrained optimization [J]. Mathematics, 2010, 19(9): 2345-2356.

[7] Akdeniz F, Yuksel G, Wan A. The moments of the operational almost unbiased ridge regression estimator[J]. Applied Mathematics and Computation, 2004, 153(3): 673-684.

[8] Arthur E, Hoerl, Robert W.Kennard. Regression:biased estimation for nonorthogonal problems[J]. Technometrics, 1970, 12(1): 55-67.

[9] Boyd S, Parikh N, Chu E, Peleato B, Eckstein J. Distributed optimization and statistical learning via the alternating direction method of multipliers[J]. Foundations and Trends in Machine Learning, 2010, 3(1): 1-122.

[10] Bioucas-Dias J M, Figueiredo M A T. 2010 2nd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing[C]. Reykjavik, Iceland: Mathematics, 2010.

[11] Breiman L. Better subset regression using the nonnegative garrote[J]. Technometrics, 1995, 37(4): 373-384.

[12] Dupe F X, Fadili J M, Starck J L. A proximal iteration for deconvolving Poisson noisy images using sparse representations[J]. IEEE Transactions on Image Processing, 2009, 18(2): 310-321.

[13] Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression[J]. The Annals of statistics, 2004, 32(2): 407-451.

[14] Elad M, Matalon B, Zibulevsky M. Image denoising with shrinkage and redundant representations[J]. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006, 2: 1924-1931.

[15] Figueiredo M A, Nowak R D, An EM algorithm for wavelet-based image restoration[J]. IEEE Transactions on Image Processing, 2003, 12(8): 906-916.

[16] Friedman J, Hastie T, Hfling H, Tibshirani R. Pathwise coordinate optimization[J]. The Annals of Applied Statistics, 2007, 1(2): 302-332.

[17] Geng B, Li Y X, Tao D C, Wang M, Zha Z J, Xu C. Parallel lasso for large-scale video concept detection[J]. IEEE Transactions on Multimedia, 2012, 14(1): 55-65.

[18] Goldstein T, O'Donoghue B, Setzer S, Baraniuk R. Fast alternating direction optimization methods[J]. SIAM Journal on Imaging Sciences, 2014, 7(3): 1588-1623

[19] He L, Tao D, Li X, Gao X. Sparse representation for blind image quality assessment[J]. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2012: 1146-1153.

[20] Hastie T, Tibshirani R, Fedman J, Franklin J. The elements of statistical learning: data mining, inference and prediction[J]. Mathematical Intelligencer, 2005, 27(2): 83-85.

[21] Hoefling, Holger. A path algorithm for the fused lasso signal approximator[J]. J. Comput. Graph. Statist. 2010, 19(4): 984-1006.

[22] Knight K, Fu W. Asymptotics for lasso-type estimators[J]. The Annals of Statistics, 2000, 28(5): 1356-1378.

[23] Lawson C, Hansen R. Solving Least Squares Problems[M]. Englewood Cliffs, NJ:Prentice-Hall, 1974.

[24] Negahban S, Martin J, Wainwright. Restricted strong convexity and weighted matrix completion: optimal bounds with noise[J]. Journal of Machine Learning Research, 2010, 13(1): 1665-1697.

[25] Nesterov Y. Introductory Lectures on Convex Optimization[M]. USA: Springer, 2004.

[26] Nesterov Y. Smooth minimization of non-smooth functions[J]. Mathematical Programming, 2005, 103 (1): 127-152.

[27] Nesterov Y. Efficiency of coordinate descent methods on huge-scale optimization problems[J]. SIAM Journal on Optimization, 2012, 22(2): 341-362.

[28] Nicholas A, Johnson. A dynamic programming algorithm for the fused lasso and L_0-Segmentation[J]. Journal of Computational and Graphical Statistics, 2013, 22(2): 240-260.

[29] Plaza A, Du Q, Bioucas-Dias J M, Jia X, Kruse F A. Foreword to the special issue on spectral unmixing of remotely sensed data[J]. IEEE Transactions on Geoscience and Remote Sensing, 2011, 49(11): 4103-4105.

[30] Richards J A, Jia X P. Remote Sensing Digital Image Analysis[M]. New York: Springer, 1999.

[31] Robert C, Peter G C, Andreas H, Gitta K, Ali P. Sparse fusion frames: existence and construction[J]. Advances in Computational Mathematics, 2011, 35(1): 1-31.

[32] Saha A, Tewari A. On the finite time convergence of cyclic coordinate descent methods[J]. SIAM Journal of Optimization, 2013, 23(1): 576-601.

[33] She Y. Sparse regression with exact clustering[J]. Electronic Journal of Statistics, 2010, 4: 1055-1096.

[34] Suzuki T. Stochastic dual coordinate ascent with altemating direction multiplier method[EB/OL].[2014-6-21]. http: //arxiv. org/.

[35] Sun Y B, Liu Q, Tang J H, Tao D C. Learning discriminative dictionary for group sparse representation[J]. IEEE Transactions on Image Processing, 2014, 23(9): 3816-3826.

[36] Tibshirani R. Regression shrinkage and selection via the lasso[J]. Journal of the Royal Statistical Society. Series B, 1996, 15(1): 267-288.

[37] Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K. Sparsity and smoothness via the fused lasso[J]. Journal of the Royal Statistical Society. Series B, 2005, 67(1): 91-108.

[38] Tibshirani R, Taylor J. The solution path of the generalized lasso[J]. The Annals of Statistics, 2011, 39(3): 1335-1371.

[39] Tibshirani R. The lasso problem and uniqueness[J]. Electronic Journal of Statistics, 2013, 7: 1456-1690.

[40] Wu T T, Lange K. Coordinate descent algorithm for lasso penalized regression[J]. The Annals of Applied Statistics, 2008, 2(1): 224-244.

[41] Yu J, Rui Y, Tao D. Click prediction for web image reranking using multimodal sparse coding[J]. IEEE Transactions on Image Processing, 2014, 23(5): 2019-2032.

[42] Yuan M, Lin Y. Model selection and estimation in regression with grouped variables[J]. Journal of the Royal Statistical Society. Series B, 2006a, 68(1): 49-67.

[43] Yuan M, Lin Y. Model selection and estimation in the Gaussian graphical model[J]. Biometrika, 2006b, 94(1): 19-35.

[44] Yuan M, Lin Y. On the non-negative garrotte estimator[J]. Journal of the Royal Statistical Society. Series B, 2006c, 69(2): 143-161.

[45] Yuan X, Alternating direction methods for sparse covariance selection[J]. Journal of scientific computing. 2012, 51(z): 26l-273.

[46] Zou H, Hastie T. Regularization and variable selection via the elastic net[J]. Journal of the Royal Statistical Sco B, 2005, 67(2): 301-320.

[47] Zhao P, Yu B. On model selection consistency of Lasso[J], Journal of Machine Learning Research, 2006, 7: 2541-2563.

[48] Zhou T, Tao D. Multitask copula by sparse graph regression[C]//Proceedings of the 20th ACM SIGKDD International conference on Knowledge Discovery and Data Mining, August 24-27, 2014, New York, USA: ACM, c2014: 771-780.

[49] Zhou T, Tao D. Shifted subspaces tracking on sparse outlier for motion segmentation[C]//Proceedings of the Twenty-Third International Joint conference on ArtificialIntelligence, August 03-09, 2013, Beijing, China: AAAI, c2013: 1946-1952.

[50] Zhu X F, Huang Z, Cui J. Video-to-shot tag propagation by graph sparse group lasso[J]. IEEE Transactions on Multimedia, 2013, 15(3): 633-646.

致　谢

在论文完成之际, 我想感谢硕士研究生阶段所有帮助和支持我的老师、同学和亲友们.

首先, 我要感谢我的导师王伟刚老师. 无论在科研还是工作生活中, 王老师都给了我很大的指导和鼓励. 读研以来, 我前进的每一步背后都倾注了王老师大量的心血. 王老师学识渊博、功底深厚, 对于科学问题总有独到的见解, 本文的不少创新点正是在与老师的讨论当中诞生的. 尽管王老师平日里要负责学院里的工作和繁重的教学任务, 但他对于科研的认真态度从不懈怠, 对于学生的学习和生活也时常放在心上. 非常幸运能在硕士研究生阶段跟着王老师学习, 师恩难忘, 在此向王老师表示最诚挚的感谢!

其次, 我要感谢讨论班的明瑞星老师、董雪梅老师. 明老师的科研态度严谨认真, 能够全面深刻地看待问题, 在课上、讨论班上总能提出创新、专业的观点, 明老师对于本论文的选题和创新点也提供了很大的帮助; 董老师为人和蔼, 经常在讨论班对学生的疑问作出透彻的解释, 给出很多相应的解决方案. 未来的学习和工作中, 我将以老师们为榜样不断前进.

此外, 我要感谢在讨论班一起学习的许耿鑫学长以及各位同学. 许学长对我论文写作过程中的指导, 让我获益匪浅. 也要感谢班里的同学,在平时一起学习、打球、健身, 使我的生活充满了乐趣.

最后, 感谢我家人、朋友、亲戚, 在我读研期间给我的生活提供了各方面的关怀, 使我保持良好的状态投入到学业中.

再次感谢所有帮助我的人们!

独 创 性 声 明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含本人为获得浙江工商大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名：　　导师签名：

签字日期：　　年　　月　　日　　签字日期：　　年　　月　　日

关于论文使用授权的说明

本学位论文作者完全了解浙江工商大学有关保留、使用学位论文的规定：浙江工商大学有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅，可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文，并且本人电子文档的内容和纸质论文的内容相一致。

本论文提交☑即日起/☐半年/☐一年以后，同意发布。

“内部”学位论文在解密后也遵守此规定。

学位论文作者签名：　　导师签名：

签字日期：　　年　　月　　日　　签字日期：　　年　　月　　日