中国人民大学

RENMIN UNIVERSITY OF CHINA

硕士学位论文

THESIS OF MASTER DEGREE

论文题目：基于强分层约束的稳健变量选择

（英文）：Robust variable selection via strong heredity constraint

作　　者：冯　倩

指导教师：易丹辉教授

2017 年 4 月 20 日

中国人民大学

硕士学位论文

（中文题目）基于强分层约束的稳健变量选择

（英文题目）Robust variable selection via strong heredity constraint

作　者　学　号：2014102955
作　者　姓　名：冯　倩
所　在　学　院：统 计 学 院
专　业　名　称：流行病与卫生统计学
导　师　姓　名：易 丹 辉

论 文 主 题 词：变量选择；密度势差异；稳健性
论文 提交 日期：2017 年 4 月 20 日

摘 要

当样本量大于变量的维数时，相较于普通最小二乘方法，最小化密度势差异（density power divergence ，DPD）是一种比较有效的稳健回归方法。借助于自适应 LASSO 惩罚函数，本文提出了 DPD-ADPLASSO 的方法，来解决带有交互效应的普通线性模型，关于变量选择和参数估计的问题，并保证强分层结构成立。相比于普通的 OLS-LASSO，通过选择更优的损失函数 DPD 来替换 OLS 和理论性质更优的 ADPLASSO 替换 LASSO，估计结果会更加稳健。文中，借助调整参数 t 惩罚参数 λ 的选取，提出了 DPD-ADPLASSO 的稳健变量选择在满足强分层约束下的基本算法，并且，在正则条件下，从理论上证明了 DPD-ADPLASSO 变量选择具有$\sqrt{n}$一致性。接着为了验证该方法的实际效果，本文通过模拟讨论了此方法在五种不同离群点类型，三种离群比例和三种不同的样本量下，DPD-ADPLASSO 稳健变量选择的表现。模拟表明：在模型误差预测方面，本文提出的 DPD-ADPLASSO 方法的效果远远优于 LAD-ADPLASSO 方法：在灵敏度以及特异度方面，当解释变量存在相关结构且样本量较小时，该方法与 LAD-ADPLASSO 相差不大，但随着样本量增大，灵敏度和特异度都比 LAD-ADPLASSO 方法效果要好，并且，总体的变量选择准确度在任何一种离群点情形下也都要优于 LAD-ADPLASSO 方法，最后，当解释变量的相关结构越来越强时，模型的估计和预测准确度会优于不分层结构下的结果。文章的最后，将 DPD-ADPLASSO 方法应用到一个满足强分层结构和强分层结构不明显的实际数据集，即血浆中 β-胡萝卜素含量和红酒质量的影响因素分析，实际数据分析表明，DPD-ADPLASSO 相比于 LAD-ADPLASSO 和普通 MM 稳健估计方法，预测准确度要明显高于 LAD-ADPLASSO 和普通 MM 方法。

关键词：交互效应；强分层；密度势差异；自适应LASSO； Oracle 性质；稳健估计

Abstract

The minimum density power divergence method （abbreviated as DPD) is a useful method for robust regression compared with ordinary least square method. Borrowing the idea of adaptive penalty, we perform variable selection on the basis of DPD. That is to say, we combine DPD and the adaptive lasso together to produce DPD-ADPLASSO method in linear model with interactional terms, at the same time, completing the variable selection with the strong heredity constraint. Compared with OLS-LASSO, DPD-ADPLASSO can do parameter estimation robustly, after replacing DPD with OLS, ADPLASSO with LASSO. Furthermore, under defined regularity conditions, our estimator enjoys $\sqrt{n}$ consistency. In order to evaluate the effect of DPD-ADPLASSO method, we carry out simulation studies under different type outliers, different outlier rates and different sample sizes. The simulation studies show that compared with LAD-ADPLASSO, DPD-ADPLASSO have smaller model error and mean square error, regarding the sensitivity and specificity, the performance of DPD-ADPLASSO and LAD-ADPLASSO have little difference when the sample size is small and there are some correlated structures in explain variables , with the sample size, DPD-ADPLASSO is better than LAD-LASSO under most cases. In addition, the overall accuracy is also better than the other method. At the end of this paper, we analyze the datasets of β Beta-carotene in plasma and wine quality from UCI. The real data analyses show that DPD-ADPLASSO method has nice performance in term of models sparsity and prediction.

Key Words : Interactional effects ; strong hierarchy; DPD; Adaptive lasso; Oracle; Robust estimation.

目 录

表格与插图

第一章 绪论

1.1 研究背景与意义

在流行病学研究中，通常一种疾病发生的危险程度，除了受到相关基因和环境特征的单独影响（即主效应）之外，还可能会受到某些基因与基因之间或者基因与环境之间的交互作用（即交互效应，又称效应修饰）的影响。研究基因与环境特征以及基因与基因之间和基因与环境之间的交互效应对疾病发生风险的影响的问题，就称为全基因组关联研究（Genome-Wide Association Studies，GWAS）。GWAS 自从被提出之后，就受到了研究者的广泛关注。

对于复杂疾病的主效应识别方面，GWAS 所识别的的易感位点仅能解释不足十分之一的遗传变异（Yajima I，2012），因此早在 2005 年，Marchini J 等就发文指出复杂疾病 GWAS 中需要研究者关注交互作用。以位列全部恶性肿瘤死因首位的肺癌为例，肺癌的发生是遗传变异和环境暴露协同作用的结局（Shields PG ，2002)。尽管肺癌主要归咎于烟草因素的暴露，烟草中含有较多的致癌物质，会导致遗传物质损伤，但是不可否认，只有部分吸烟者发展为肺癌，并且非吸烟者中也有肺癌，这些事实说明，个体的差异对环境暴露的反响存在遗传易感性。此外，大量的研究表明在吸烟暴露相似的情形下，具有肺癌家族史的个体具有更高的患肺癌风险，进一步说明了遗传易感因素在肺癌发生发展中的重要性。如果借助识别出的基因与家族史的交互作用，便有利于识别出肺癌的高危人群，进而尽早实施控烟干预，未雨绸缪，防患于未然。再比如，苯丙酮尿症只有在遗传变异和环境的暴露同时存在时，症状才会发生，在这里遗传变异是指苯丙氨酸经化酶缺乏，环境的暴露是饮食中的苯丙氨酸。因此，在疾病影响因素的相关研究中，考虑基因与环境特征之间的交互效应是具有重要意义的，一方面能够相对准确的估计出复杂疾病的遗传和环境效应，使得研究结果

更加充分可靠；另一方面还会发现一些有价值的开创性知识，为疾病的预防和治疗提出全新的办法。

在一些相关研究中分层结构（hierarchical structure）也称边际准则（marginality principle，Nelder, 1994; McCullagh, 2002），或遗传性（heredity，Chipman, 1996; Chipman, Hamada & Wu, 1997）。当协变量之间存在交互效应时，一般认为主效应和交互效应之间应当自然而然地满足一种分层结构，也即主效应为第一层（又称父层），交互效应为第二层（又称子层），模型中要想存在交互效应，就必须存在相应的主效应。也就是说存在某个交互效应的前提是存在各自相对应的主效应。

之所以要求交互效应满足主要是基于实际应用方面的考虑。在这里，需要介绍两个概念，参数稀疏性（parameter sparsity）和实际稀疏性（practical sparsity）（Bien, Taylor & Tibshirani, 2013）。参数稀疏性是指统计学领域，统计学家在做基于回归模型的变量选择问题时，模型包含的重要协变量个数，也即非零协变量的个数。然而实际稀疏性是针对在实际应用中，那些做调查的人真真正关心的。当模型不满足不分层结构时，也即模型中存在交互效应，但不存在相应的主效应，这些调查人员的劳动量并未减小，也即实际稀疏性是指按照未来做预测的目的而需要测量的解释变量个数。举例说明：如果有两个相同参数稀疏性为 K 的模型，但一个满足分层结构，则实际稀疏性仍为 K，但当另一个不满足结构时，实际稀疏性就变成 K!级别。因此通常在研究交互效应模型时，研究分层结构不仅仅兼顾了统计学家关心的参数稀疏性，也更受实际调查工作者的欢迎和青睐。同时，要求交互效应满足分层结构的拎一个主要原因，实际上也是做到尽可能对已测量变量信息的充分利用，使得用尽可能比较少测量的解释变量来对响应变量做解释和预测。

随着基因测序技术的发展和各种测量手段的快速进步，实际应用中，在研究个体患某种疾病的影响因素时，通常可以测量得到大量基因与环境等有关变量的信息。然而，在这大量的基因与环境因素中，真正对所研究疾病有重要影响的因素通常是少数的。当纳入众多无关协变量进行统计建模和预测时，将使得模型特别复杂，解释性差，且容易引起协变量之间的多重共线性，造成模型预测精度低，统计推断失效。另外，众多无关的协变量也很有可能湮没了那些真正起重要作用的协变量，造成对实际现象解释和本质规律认识的偏误甚至背

离，这在临床医学研究等实际问题中可能会造成严重后果。因此，在生物医学或流行病学研究中经常遇到的一个问题就是疾病危险因素的筛选，也即从大量可测的基因或环境因素中识别出某种疾病发生的真正危险因素，而剔除掉那些对该疾病的发生影响甚微以至可以忽略的因素，这个过程就是变量选择（variable selection）。

变量选择就是对协变量进行降维以达到协变量稀疏性的过程，有效的变量选择作用主要有如下两个方面：一是通过踢除多余的协变量，选择最优的解释变量，进而精简模型，加强模型解释性；二是避免多重共线性问题，使得参数估计更加稳定，提高模型整体预测精度。变量选择方法主要精力来两个主要的阶段，第一阶段为传统的最优子集变量选择（subset selection），即先选择 AIC（Akaike, 1973）、BIC（Schwarz, 1978）、Cp（Mallows, 1973）、CIC（Tibshirani & Knight,1999）等信息准则的基础之上，再选择一个较好的拟合途径，如逐步回归，向前选择等。但该阶段的方法存在一个缺点，也即当变量个数较大时，此方法等计算量偏大，耗时严重。后来随着研究的深入，统计学者们提出了基于惩罚函数的变量选择方法，并由此进入变量选择方法的第二个阶段。这类方法的思想核心是：之前的目标函数如最小二乘 (OLS) 或最大似然 (MLE) ，在它们的基础上，再加入惩罚函数，构成新的目标函数，然后最小化或最大化此目标函数得到参数的估计值。这类方法之所以能实现变量选择，主要原因在于引入了惩罚函数，惩罚函数中含有调整参数，通过惩罚函数调整参数的选取，将一些不显著的变量的回归系数压缩为零，而将显著的变量系数压缩很小。随着研究的不断深入，压缩强度也引入了权重来调节。近年来迅猛发展的诸如 nonnegative garrote（Breiman, 1995）、Lasso（Tibshirani, 1996）、Bridge（Knight & Fu, 2000）、SCAD（Fan & Li, 2001）、LARS（Efron et al., 2004）、Elastic Net（Zou & Hastie, 2005）、Adaptive Lasso（Zou, 2006）、Group Lasso（Yuan & Lin, 2006）、MCP（Zhang, 2010）等的收缩惩罚（shrinkage penalization）方法，其中 Fan & Li 提出了变量选择方法应当满足的三条性质，即 Oracle 性质。

（1）稀疏性：参数估计应该自动实现系数的稀疏性，也即自动的将一些不是那么重要的解释变量系数压缩到零。

（2）无偏性：参数估计无偏或近似无偏。

（3）连续性：为保证模型预测的稳定性，参数估计对于数据连续。

SCAD 和 Adaptive Lasso 已被证明具有上述的 Oracle 性质。总而言之，变量选择方法的理论与应用研究在近年来得到了极大的关注与发展。

对交互效应模型做变量选择并保证模型分层结构的成立称为分层变量选择（hierarchical variable selection）。但是，不可否认，随着交互效应的引入，不但会增加变量维数，更会使得模型结构复杂化，给模型参数的估计和统计推断带来更大的困难与挑战。常规的 Lasso 等方法将所有的协变量（包括主效应和交互效应）独立对待，因此变量选择结果通常不满足分层结构。此外，关于变量选择过程中的统计推断，在回归分析里，参数的显著性检验和置信区间的估计，对辅助评价模型和鉴别噪声变量具有显著的意义。然而，在交互效应模型的分层变量选择中，研究尚较为欠缺，也没有普遍的认可的方法和准则。因此，对于交互效应模型的分层变量选择问题，除了变量选择方法本身的难点之外，同时还面临着一个巨大挑战：

即数据存在离群点的问题。线性回归模型中，研究者一般假定误差项服从一定的条件，比如独立同分布于正态分布或方差有限。此时最小二乘估计以及极大似然估计的估计结果具有最佳线性无偏性，可是实际应用中，数据很容易受到离群点的影响，导致误差项并不是严格正态的。例如章节 5.1 血浆中胡萝卜素的数据，从其响应变量的直方图以及 QQ 图 (5.1) 可以看出，该数据存在离群点，不是严格正态的。已经有研究证明，即使很少量的离群点也会给模型估计系数带来严重的偏差。因此，在这种情形下，采取稳健的估计方法来实现参数估计和变量选择就尤其必要。但是目前这方面的相关研究也十分缺乏。

1.2 相关文献综述

由于本文的主要内容就是在对带有交互效应模型的基础上变量选择，并且满足强分层结构和实现更为稳健的估计。由于稳健估计是由损失函数 ，而罚函数则实现了变量选择的目的。因此本节从以下三个方面，对本研究（1）目前已有的稳健变量选择方法研究（2）满足强分层结构的带有交互效应模型的变量选择。主要从普通线性模型和带有交互效应模型的变量选择已有方法这两方面来系统阐述（3）进一步对此基础模型做统计推断的三个相关子问题概述其研究现

状，并进行全面的文献综述。

1.2.1 稳健变量选择方法研究

在线性回归模型里，研究者一般会首先假定，模型需要误差项满足一定的条件，独立同分布于正态分布或方差有限，通常情况下选择最小二乘估计是比较经典的估计。但是在实际数据中，数据会有很大可能出现离群点，有可能出现在解释变量，也有可能来源于响应变量。已经有研究者发现，当数据稍微改变一点点，其参数估计结果就会发生巨大的变化 (Seber & Lee, 2003)。一方面，线性回归模型中忽略显著的解释变量会导致有偏的参数估计和预测，包含了不显著的变量会导致参数估计的效率降低以及不精确的预测。另一方面，OLS 以及 MLE 受离群点的影响比较大，此时基于 OLS 以及 MLE 的变量选择方法同样不稳健。因此对于存在离群点的数据，寻找一种对离群点不是特别敏感的方法来替代普通的最小二乘方法，就显得尤为必要。自 1960 年以来，许多稳健的回归方法被提出来。

Huber (1964)提出了 M 估计（Maximum Likelihood Type Estimates, M-Estimates）稳健回归方法, 也称作广义最大似然估计。M 估计具有一系列良好的性质，比如当目标函数为凸函数，并且满足一些相对弱的条件时，M 估计具有非常高的效率，弱相合性和渐近正态性。但是 M 估计也有一个缺点：当离群点是来自响应变量时，该方法比较稳健，但是对于离群点来自解释变量，表现就不够稳健，甚至该方法还不会比最小二乘估计要好。

为了克服 M 稳健回归方法不够稳健的缺点，Rousseeuw & Yohai (1984) 提出了 S 估计。S 估计的提出要是基于最小化残差尺度而得到的。S 稳健回归克服了 M 估计对离群点来自解释变量的情形不够稳健的缺点，也具有渐进正态的良好性质。但是该方法不是特别有效；

为了结合 S 估计的稳健性以及 M 估计的有效性，Yohai & Zamar (1988) 提出了 MM 估计；它集中了 S 估计的高崩溃点（Breaking Point, BP）和 M 估计的高效率的优点。MM 估计的步骤是第一步先使用 S 估计，得到回归系数的初始估计值，然后选用效率标高的 Turkey 双权型函数进行 M 估计。

当然，除了上述所说的线性稳健回归等诸多办法，还有基于秩次的 R 类稳

健回归等。一般的，研究线性稳健回归在实际问题的应用时，通常将方法进行改进或多种稳健方法结合起来使用。

近年来，稳健回归方法因其巨大的实际意义受到越来越多统计学者的重视，Wang, Li & Jiang (2007) 提出了 the least absolute deviation (LAD)的损失函数，它的形式为：

$$\Phi(y_i - x_i^T\beta) = \left|y_i - x_i^T\beta\right| \tag{1.1}$$

罚函数为普通的 LASSO 罚。因此，总的来说，总体的目标函数是：

$$\sum_{i=1}^{n}\left|y_i - x_i^T\beta\right| + n\sum_{j=1}^{p}\lambda_j\left|\beta_j\right| \tag{1.2}$$

其中惩罚参数λ采用类似 BIC 的准则选取，即：

$$\lambda_{nj} = \frac{log(n)}{n\left|\beta_j^{LAD}\right|} \tag{1.3}$$

其中，β_j^{LAD}是 LAD 的估计结果。文章中证明了当参数按照上述准则选取时，LAD-LASSO 估计结果具有$\sqrt{n}$一致性和 Oracle 性质。并且文章的理论和模拟表明，在回归模型中，当响应变量来自于厚尾分布时，LAD-LASSO 可以得到一致的变量选择效果，并且计算量相对于 BIC、AIC 准则较小。但是文中只考虑了误差项非正态且为厚尾分布，并没有考虑解释变量存在离群点时的情形。

Leng (2010) 研究了通过引入正则化秩回归估计量实现稳健估计。该方法克服了 LAD 的在某些情况下效率较低的缺点，并且结合 OLS 以及 LAD 的优点，提出此新的稳健变量选择方法，可以获得较高的效率和稳健性。该方法的步骤是：先对数据进行转换，即将原始数据的差值作为新的数据，对这个新的数据建立一个 LAD 模型，得到回归模型系数的一个初始估计值。然后第二步就是采用自适应 Lasso 作为罚函数的基础上，最小正则化最小二乘得到最终的估计值，文章中采用 LARS 算法进行估计。即目标函数为最小化下式：

$$\sum_{i=1}^{n}\left(y_i - x_i^T\beta\right)^2 + n\lambda_n\sum_{j=1}^{p}w_j\left|\beta_j\right| \tag{1.4}$$

惩罚参数λ通过得分信息准则(SIC) 进行选取。并且利用该准则选取出的惩罚参数，作者已经证明该方法具有 Oracle 性质。在模拟章节，可以发现当离群点来自响应变量，离群点比例为 0.10 时，本方法无论是变量选择效果还是模型

误差方面都要比普通自适应 LASSO，LAD-LASSO 要好。文章中的模拟只研究了离群点来自响应变量且固定的离群点比例为 0.10 时该方法的效果，没有研究离群点来自解释变量和不同离群点比例的情形。

Wang, J & Zhang (2013) 提出了指数平方损失函数下的稳健变量选择。该篇文章从理论上刻画了稳健性。其损失函数为：

$$\Phi(y_i - x_i^T\beta) = 1 - \exp(\frac{(y_i - x_i^T\beta)^2}{\gamma}) \tag{1.5}$$

其中 γ 是调节稳健性的指标。显然, 如果 γ 取值很大，其损失函数可以近似为最小二乘, 如果 γ 取值较小，那么相比于最小二乘，残差较大的项对于目标函数的影响较小，从而更加稳健。罚函数为自适应 LASSO 。即总体的目标函数是最小化下式：

$$\sum_{i=1}^{n} 1 - \exp\left(\frac{(y_i - x_i^T\beta)^2}{\gamma}\right) + n\lambda_n \sum_{j=1}^{p} w_j|\beta_j| \tag{1.6}$$

稳健性调节参数 γ 通过最小化渐进方差的行列式得到。w_j为 MM 估计的倒数。惩罚参数λ 通过类似 BIC 准则进行选取。并且利用该准则进行选取后，作者证明了此种参数估计方法具有一致性和 Oracle 性质。在数据模拟章节，作者相对充分地考虑了误差项来自于解释变量，以及来自响应变量的情况。模拟表明，该方法的变量选择效果和模型误差相对于普通的自适应 LASSO，LAD-LASSO，CQR-LASSO 效果要好，并且实际数据的应用时，作者提出的方法更倾向于选择出一个更加稀疏的模型。但是文章有一个缺点，并没有考虑随着数据中离群点比例的变化，该方法的表现效果如何。

Basu et al., (1998) 提出了一种稳健并且有效的参数估计准则：最小化密度势差异(density power divergence，DPD) 。该方法基本思想是首先定义生成数据的分布，即非参数分布族 G 和我们假定的分布 F（参数分布族） 所对应的密度函数 g 和 f 之间的差异，然后通过最小化该密度势差异得到参数的估计值。DPD 方法利用经验分布函数来代替生成数据的真实分布，这样避免了非参数方法估计数据真实分布时带来的带宽的选择问题，使方法更为清晰易懂。

对如下参数分布族$\{F_\theta\}$，f_θ是对应的密度函数，其中θ是未知参数。定义参数分布族{G}，对应的密度函数是 g。定义f_θ与 g 之间的密度势差异：

$$d_t(g,f)=\int\left(f_\theta^{1+t}(x)-\frac{t+1}{t}g(x)f_\theta^t(x)+\frac{1}{t}g^t(x)\right)dx \tag{1.7}$$

DPD 方法中包含调节调节稳健性和有效性的指标 t。t 越大越稳健但是有效性越低，反之越不稳健有效性越高。理论上其调节参数 t 可以取所有非负数。但是根据研究(Basu et al., 1998)显示，调节参数取值大于 1 时，有效性降低会很多，所以实际应用时只考虑 t 取值为 (0, 1]。由于数据中离群点的存在，将数据的真实分布 G 用经验分布代替，那么公式（1.7）即转化为最小化下式：

$$\int f_\theta^{1+t}(x)dx-\frac{t+1}{t}\frac{1}{n}\sum_{i=1}^{n}f_\theta^t(x_i) \tag{1.8}$$

Durio & Isaia (2011), Ghosh & Basu (2013) 将 DPD 稳健的参数估计方法应用到线性模型中。线性回归模型中对于每一个样本观测，可以得到一个密度势差异，通过最小化所有样本平均的密度势差异，得到回归系数的估计值。并且证明了，当生成数据的真实分布为正态分布时，其估计量具有一致性和渐进正态性，Durio & Isaia (2011) 还定义了模型间相似性标准化指数来选取最优的调节参数 t。在回归模型中，每一个分布$\mathrm{G_i}$同样用经验分布函数代替，但是每一个分布只有一个观测，因此经验分布变为退化分布。通过公式（1.20），可以得到每一个观测的 DPD，结合正态分布$N(x_i{}^T\beta,\sigma^2)$的密度函数，最终的目标函数是最小化：

$$\frac{1}{(2\pi)^{t/2}\sigma^t\sqrt{1+t}}-\frac{t+1}{t}\frac{1}{(2\pi)^{t/2}\sigma^t}e^{-t\left(y_i-x_i{}^T\beta\right)^2/2\sigma^2} \tag{1.9}$$

对于线性回归模型，上述的稳健回归方法很少讨论离群点来自解释变量，或者离群点既有来自解释变量又有来自响应变量的情形。基于 DPD 方法的变量选择目前也没有很多在上面这些情形的研究。本文将研究以DPD作为损失函数，自适应 LASSO 作为罚函数，对带有交互效应的线性模型进行稳健变量选择，还满足强分层的约束。为响应变量服从正态分布，离群点来自解释变量的数据，或者离群点来自响应变量 ，也即导致误差项是一个混合分布，其中绝大多数来自正态分布的数据，或者离群点不仅有来自解释变量，还有来自响应变量的数据，提供一个同时进行变量选择和参数估计的新途径。

1.2.2 带有交互效应模型的变量选择

由于本文的主要思路是将稳健变量方法应用在带有交互效应的普通线性模型中。因此有必要从带有交互效应的普通线性模型的变量选择角度来展开详细讨论。接下来将分别从普通线性模型和带有交互效应的普通线性模型下之变量选择方法展开文献综述，也即阐述本文所研究的基本模型在国内外的现有理论方法。

对于回归模型的变量选择问题，惩罚因子法（又称收缩惩罚法）是目前比较受到研究者青睐的主要方法。考虑如下的简单线性回归模型

$$y = \sum_{j=1}^{p} x_j \beta_j + \varepsilon \tag{1.10}$$

其中$y = (y_1, \dots, y_n)'$为响应变量，$x_j = (x_{1j}, \dots, x_{nj})'$为解释变量，$\beta = (\beta_1, \dots, \beta_p)'$为回归系数向量，误差项$\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)'$相互独立，且$\varepsilon_i \sim N(0, \sigma^2)$，$j = 1, \dots, p$，$i = 1, \dots, n$，惩罚因子法是在普通最小二乘（ordinary least squares，OLS）作为损失函数的基础上，加入惩罚函数，从而对回归系数β_j施加压缩，得到如下惩罚最小二乘问题

$$\min_{\beta} \frac{1}{2n} \left\| y - \sum_{j=1}^{p} x_j \beta_j \right\|^2 + P_\lambda(|\beta_j|) \tag{1.11}$$

其中$\|v\| = \|v\|_2 = (v'v)^{1/2}$为$L_2$范数，$P_\lambda(|\beta_j|)$为惩罚项，含有一个惩罚参数$\lambda$（$\lambda > 0$），惩罚项的作用是对回归系数$\beta_j$进行压缩，某些系数就可能被压缩为0，从而达成变量选择的目标。

惩罚项的一个常用形式是$P_\lambda(|\beta_j|) = \lambda \sum_{j=1}^{p} |\beta_j|^q$,其中$q \geq 0$。可以发现当$0 \leq q \leq 1$，由于惩罚项在零点不可导，也称存在奇异点，就会以相对更大的概率将一些不重要的变量系数压缩为零，从而该方法就可以实现估计出的部分解释变量系数为零，从而实现变量选择。当$q = 1$时，就是经典的 Lasso（Least absolute shrinkage and selection operator）惩罚项，该方法于 1996 年由 Tibshirani 提出，而且当设计矩阵正交，该方法可以有显式解。当$q \geq 1$时，就是桥方法（Bridge），特殊的，当$q = 2$时，就是普通的岭回归估计。Lasso 方法相比于岭回归估计，它的优点在于对于一些参数估计系数较大的变量压缩较轻，从而提

高了参数估计的精确性。不过，虽然 Lasso 办法的参数估计具有连续性，其后 Fan 和 Li（2001）证实出 Lasso 方法不具有 Oracle 性质；并且它属于预测准确度较高，可惜相合性较差。

随后，针对 Lasso 方法的缺点，该方法得到了进一步的修正，Zou 和 Yuan（2006）提出了自适应 Lasso（Adaptive Lasso）。该方法的核心是对不同的变量施加不同程度的罚，也即根据全模型下最小二乘的估计结果，如果估计出的系数较大，则施加小一点的罚，而一旦估计出的系数值较小，则施加比较大的罚，以致压缩到零。注意到系数的权重全是根据从数据中提取的，所以是自适应 Lasso，自适应 Lasso 还被证明具有 Oracle 性质，此外，当变量个数固定，样本量趋于无穷式，方法还具有较好的相合性（见李根，2012）。因此本文在后续的讨论中，将使用自适应 Lasso 作为最终的罚函数。当然，还有 SCAD、MCP、Group Lasso、Fused Lasso，弹性网惩罚等其他不同的惩罚函数被先后提出，不同形式的罚函数也自然而然对应着不同的变量选择方法。

以 Lasso 为代表的不同的罚函数，已被广泛应用在简单线性回归模型的变量选择问题里，然而对于如下线性交互效应模型

$$y = \sum_{j=1}^{p} x_j \beta_j + \sum_{j=1}^{p-1} \sum_{j'=j+1}^{p} x_{jj'}^{*} \gamma_{jj'} + \varepsilon \tag{1.12}$$

其中$x_{jj'}^{*} = \left(x_{1j}x_{1j'}, \dots, x_{nj}x_{nj'}\right)'$为协变量$x_j$和$x_{j'}$的交互效应，$j < j', j, j' = 1, \dots, p$，直接使用普通的 Lasso 等罚函数，会将主效应β_j和交互效应$\gamma_{jj'}$同等对待，也即不能满足主效应β_j和交互效应$\gamma_{jj'}$之间的分层结构。

交互效应模型中的分层结构包括两种形式，强分层（strong hierarchy）和弱分层（week hierarchy），前者要求只要有一个主效应系数为零，那么对应的交互效应系数也必然为零，等价地，交互效应存在的前提是相应的两个主效应均存在，而后者要求交互效应存在的前提是至少有一个相应的主效应存在，也即两个主效应系数均为零，交互效应系数才为零。对上述的线性交互效应模型（1.12），强分层结构意味着

$$\gamma_{jj'} \neq 0 \implies \beta_j \neq 0 \text{ 且} \beta_{j'} \neq 0, \quad \forall j, j' = 1, \dots, p \tag{1.13}$$

而弱分层结构意味着

$$\gamma_{jj'} \neq 0 \implies \beta_j \neq 0 \text{ 或} \beta_{j'} \neq 0, \quad \forall j, j' = 1, \dots, p \tag{1.14}$$

由于在实际应用中，研究者更加关注的强分层条件，所以在本文的研究中，选择更具实际价值的强分层进行研究。

对于上述模型（1.12）的带有强分层约束的变量选择问题，最小角回归算法（Least Angel Regression, LARS）因为计算简便及良好的预测能力是比较常用的办法，见 Yuan, Joseph & Lin (2007)。该算法的核心思想是先找出和响应变量相关度最高的那个变量，然后逐步的调整它与残差的系数，使之慢慢减小，直至相关性不是非常显著的时候，再来寻找第二个变量，与向前逐步回归(Forward Stepwise)的算法有相似的地方，但是后者是每次找出变量的子集来拟合模型。

Yuan, Joseph & Zou (2009)在 Nonnegative Garrote（NG）方法的基础上，施加全新的不等式约束来满足变量选择的分层结构。其实 Lasso 方法也是有 NG 发展而来。NG 方法的原理是不加罚函数项，只是在损失函数的$x_j\beta_j$部分前面加入满足线性约束条件的惩罚参数θ。

$$\min \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} c_j \beta_j^0 x_{ij})^2 \tag{1.15}$$

其中c_j为正的压缩系数，c_j的和有上界。

Zhao, Rocha & Yu (2009)提出 Composite Absolute Penalties（CAP）方法来对交互效应实行强分层的约束条件，其罚函数项形式为

$$\lambda \sum_{j<j'} \left(|\gamma_{jj'}| + \|(\beta_j, \beta_{j'}, \gamma_{jj'})\| \right) \tag{1.16}$$

并且此方法被用于高维数据的成组变量选择中，并且通过模拟表明，要比普通的 Group Lasso 变量选择和估计效果好。

Radchenko & James (2010)研究了非线性交互效应模型的变量选择，对于线性交互效应模型（1.3），作者首先拓展此式为：

$$y = \sum_{j=1}^{p} f(x_j) + \sum_{j=1}^{p-1} \sum_{j'=j+1}^{p} f(x_{ij}^*, x_{ij'}^*) + \varepsilon \tag{1.17}$$

并提出 Nested-group Lasso 罚来保证分层结构，Nested-group Lasso 罚函数的形式为

$$\lambda_1 \sum_{j} \|(\beta_j, \gamma_j)\| + \lambda_2 \sum_{j<j'} |\gamma_{jj'}| \tag{1.18}$$

其中$\gamma_j = (\gamma_{1j}, \cdots, \gamma_{j-1,j}, \gamma_{j,j+1}, \cdots, \gamma_{jp})'$为对应主效应系数$\beta_j$的交互效应系数。

Bien, Taylor & Tibshirani (2013)将主效应系数β_j改写，即$\beta_j = \beta_j^+ - \beta_j^-$的形式，然后利用带不等式约束的 Lasso 惩罚项

$$\lambda_1 \sum_j (\beta_j^+ + \beta_j^-) + \lambda_2 \sum_{j<j'} |\gamma_{jj'}|, \ \text{s.t.} \|\gamma_j\|_1 \le \beta_j^+ + \beta_j^-, \beta_j^+ \ge 0, \beta_j^- \ge 0 \quad (1.19)$$

其中$\|\cdot\|_1$为L_1范数。罚函数（1.19）可以转化为$\lambda \sum_j \left[\max\left\{|\beta_j|, \|\gamma_j\|_1\right\} + \|\gamma_j\|_1/2\right]$的形式，类似于 Nested-group Lasso 罚，能够以概率 1 保证变量选择结果满足强分层结构。

Lim & Hastie (2013)将线性交互效应模型改写为

$$y = \sum_{j=1}^{p} x_j \beta_j + \sum_{j=1}^{p-1} \sum_{j'=j+1}^{p} \left(x_j, x_{j'}, x_{jj'}^*\right)\left(\tilde{\beta}_j^{(j')}, \tilde{\beta}_{j'}^{(j)}, \gamma_{jj'}\right)' + \varepsilon \quad (1.20)$$

并采用如下的 Overlapped group Lasso 罚函数来保证分层结构

$$\lambda \sum_j |\beta_j| + \lambda \sum_{j<j'} \left\|\left(\tilde{\beta}_j^{(j')}, \tilde{\beta}_{j'}^{(j)}, \gamma_{jj'}\right)\right\| \quad (1.21)$$

上述研究基于模型变换、惩罚函数改进或二者结合的方法，同时对主效应和交互效应进行变量选择并保证分层结构，有的具有一致性等良好的理论性质，但其普遍的缺点是优化算法复杂、计算量巨大。

关于交互效应满足强分层的变量选择方法，另外一类实际可行的办法是两阶段选择—首先忽视交互效应，只对主效应系数进行估计，估计后固定主效应的值，利用强分层的要求，只有当模型中已包含相应的主效应时才需要考虑交互效应的选择，否则直接将相应的交互效应剔除即可（见 Park &Hastie, 2008; Wu et al., 2009; Wu et al., 2010）。尽管上述的两阶段方法具有计算上的优势，但其理论有效性扔有待考证。

然后，受上述两阶段变量选择方法思想的启发，Choi, Li & Zhu (2010)将交互效应系数$\gamma_{jj'}$写成$\gamma_{jj'} = \psi_{jj'}\beta_j\beta_{j'}$的形式，然后对$\beta_j$和$\psi_{jj'}$施加 Lasso 罚进行变量选择。通过这种参数变换，变量选择结果必然满足强分层结构，也即若$\beta_j = 0$或$\beta_{j'} = 0$，则自然地有$\gamma_{jj'} = 0$，并且证明出具有 Oracle 的理论性质。Zhu, Zhao & Ma (2014)提出 Progressive 估计方法来保证交互效应满足强分层结构，该方法的核心是在每一步迭代估计时，只有当两个相应的主效应都非零时才需要更新交

互效应，否则直接令此交互效应系数为零。另外，这三种方法最明显的优点是计算过程清晰易懂，计算量小，实际可行性强。都解决了在带有交互效应的普通线性模型下，如何实现满足强分层约束的变量选择。本文就创新性结合这两种方法，来实现主效应系数β_j和交互效应系数$\gamma_{jj'}$的估计。

但以上方法均没有考虑在线性交互效应模型（1.12）基础之上，当数据出现离群点时，如何更为稳健的估计主效应β_j和交互效应系数$\gamma_{jj'}$。

1.2.3 变量选择中的统计推断

回归分析中的两大主题是参数估计与统计推断，前者是根据样本数据求解理论回归模型，后者则对参数估计结果和模型拟合效果等，做进一步的诊断和评价，对检测模型中的噪声变量，评价所研究模型有效性以及精简回归模型都具有重要作用。在普通最小二乘回归中，参数显著性检验（构造回归系数显著性p值）和对应的置信区间估计是关于回归系数的统计推断的两大问题，相关的理论方法研究在当代统计学中也相当充分和完善。对于传统的逐步回归或子集选择等变量选择方法，回归系数的估计仍然是 OLS 估计，因此参数的统计推断与普通最小二乘所用的统计推断步骤相同。然而，对于 Lasso 等收缩惩罚变量选择问题，由于参数的惩罚估计有偏，且其分布未知，本来应用在普通最小二乘回归中的参数显著性检验、置信区间估计等传统的统计推断方法失效。

线性回归模型的惩罚变量选择中的统计推断问题因其重要的实际应用价值，近几年受到越来越多的关注和研究。Wasserman & Roeder (2009)提出样本分割（sample-splitting）方法来进行变量选择并构造回归系数的显著性检验 p 值，其基本思路是将样本数据随机均等分割成 S1 和 S2 两部分，首先使用 S1 部分的数据进行变量选择，然后基于 S2 部分的数据对选择出来的低维协变量做 OLS 估计并求回归系数的显著性 p 值，未选择出的变量 p 值为 1。该方法的缺点是参数推断的结果很容易受到样本数据的单次随机分割的影响，也即结果很不稳定。Meinshausen, Meier & Bühlmann (2009)提出改进的多重样本分割（multi sample-splitting）的方法，将样本分割过程重复 B 次，每一次都得到所有系数的显著性 p 值，然后将 B 次结果汇总计算出每个回归系数最终的显著性

p 值。

Minnier, Tian & Cai (2011)使用重抽样方法近似回归系数惩罚估计的分布，从而构造参数的假设检验过程和置信区间估计。Chatterjee & Lahiri (2011, 2013)采用 Bootstrap 方法来近似 Lasso 和 Adaptive Lasso 估计的分布并构造参数惩罚估计的置信区间。上述样本分割方法或重抽样方法的缺点是，统计推断结果一定程度上受样本数据的分割或重抽样过程的影响，具有不稳定性。另外一些研究使用全部样本数据构造参数惩罚估计的渐近分布或直接构造检验统计量来对参数做统计推断，从而避免了这种不稳定性。Bühlmann (2013)通过使用 Lasso 对岭回归（Ridge regression）估计进行偏倚调整来构造得到回归系数的渐近分布，进而对回归系数进行假设检验，并得到显著性 p 值。Zhang & Zhang（2014）提出低维映射（low dimensional projection）方法来对一组事先确定的低维回归系数构造置信区间估计。Lockhart et al. (2014) 在 Lasso 回归的连续求解路径中，根据每一条求解路径的估计结果来构造协方差检验统计量（covariance test statistic），从而对进入当前模型的协变量实行显著性检验。另外，Javanmard & Montanari（2013）和 Van De Geer et al. (2014)通过对 Lasso 估计进行偏倚调整来使其服从渐近正态分布，从而可以构造调整的 Lasso 估计的置信区间和假设检验。

1.3 本文创新点摘要

本文的研究目的是将 DPD 的稳健估计方法应用在带有交互效应的普通线性模型中，同时满足交互效应的强分层结构。首先 DPD 方法在目前已有文献中只应用在普通线性模型和广义线性模型里，虽然 Ma(2016)有将 DPD 推广到高维线性模型中，但其中的交互项与本文的基本模型不一致。因此从选题角度上，是首次创新性的将 DPD 应用在带有交互效应的普通线性模型里。

其次，当把基于 DPD 的稳健估计方法应用在此模型时，既需要考虑 DPD 算法的本身的迁移，还要考虑如何实现强分层，再者还需要得到变量选择的效果，如何将这三种思路结合起来从而解决本文提出的问题，是此篇文章的重难点所在。基于 DPD 损失函数和自适应 LASSO 罚函数的基本框架，本文作者提出了一种迭代算法的改进，该改进后的算法是在将主效应和交互效应同等对待的基础上，结合 SHIM 算法，通过分别调整 DPD 的调节参数和罚函数的惩罚参数，就可以得到该模型满足强分层约束条件下的系数估计值。

那么，从算法本身的一些创新点体现在如下两方面：

（1） 算法内部方差的估计。在块坐标梯度下降算法 (Tseng & Yun, 2009) 的基础上，为了估计方差的初始值，采用二分法来替换对初值敏感的牛顿迭代法。虽然在数值模拟中发现相差不大，但二分法相对更加稳健安全，而且算法的实现上要比牛顿迭代法容易的多。

（2） 稳健性调节参数的选取。这一类是在改进算法内部，DPD 所产生的调节参数 t。由于 DPD 方法中涉及到调节稳健性与有效性的指标，所以实际应用中需要对其进行选择。本文采取一串 t 值的方法，进而结合预测误差最小时选取最优的 t 值。

接着，为了证明提出的算法的有效性，本文除了从理论上证明出 DPD-ADPLASSO 的估计量具有$\sqrt{n}$一致性，还从数值模拟的角度来进一步说明变量选择效果的一致性。

自然而然的，接下来的一个创新点是模拟情形的多样化。本文首次比较系统的尽可能全面的考虑将 DPD 方法在应对不同离群点时的表现情况。也即从

解释变量离群，响应变量离群，解释变量和响应变量均离群三个方面，和在解释变量和响应变量都离群的情形下，解释变量的相关结构依次是相互独立，呈带状（banded）结构和方差为指数衰减情况（即解释变量相关结构逐渐增加）这些情况全方位的考察在不同离群点来源、不同的离群比例和不同的解释变量相关结构下，DPD-ADPLASSO 的模型估计和预测效果。

最后值得一提的是将同样稳健的 LAD 方法也应用此算法，也即 LAD-ADPLASSO 的算法。同时将普通的稳健 MM 方法作为基准（Oracle 方法），来对 DPD-ADPLASSO，LAD-ADPLASSO 和 Oracle 三种方法做比较。

最终实现对线性交互效应模型中的参数主效应、参数交互效应同时进行变量选择与参数估计，并保证变量选择的强分层结构成立。

1.4 本文研究内容

本文的研究框架如下：第一章，介绍本文的研究背景与意义，以及带有交互效应的普通线性模型在变量选择和稳健变量选择方法上的研究现状，主要包含密度势差异基本理论、罚函数的基本理论、坐标下降算法等；第二章，介绍了带交互效应的线性模型基本形式，并依据此模型，本文创新性的提出了 DPD-ADPLASSO 稳健变量选择方法问题，并给出相应的算法，在算法中详细阐述 DPD-ADPLASSO 稳健变量选择方法中涉及四类参数的选取，回归系数，回归方差，稳健性与有效性调节参数以及惩罚调节参数，第三章，通过模拟表明 DPD-ADPLASSO 稳健变量选择方法的一致性和有效性；第四章，将本文提出的变量选择方法应用到血浆中 β-胡萝卜素含量和红酒质量的影响因素分析，并发现了本文的方法保持了较高的预测准确度。第五章，总结和展望。对文章的研究进行总结，并提出可进一步研究的问题与思路。附录中证明了本文提出的方法的估计量具有一致性的良好理论性质。整体内容框架如下图 1.1：

总体来说，本文要解决的问题是对于含有交互效应的线性模型，一方面要实现变量选择的目的，一方面要达到稳健估计，最后需要满足交互效应和主效应间的强分层结构。因此如何将这三个问题一并解决是本文的重点所在。本文

在结合 Zhu(2010)和 Xu（2016）算法的基础上，做了合并，并做出了相应的算法改进，比较合适的解决了上述问题。本文在从理论性质上证明出了该方法下的估计量具有一致性，还利用具体的数据模拟来进一步从量化的角度去探讨，该方法在不同样本量，离群比例，离群来源等的模型估计和预测效果。因此接来的第二章介绍了模型的基础形式和主要算法，第三章详细设置了模拟的五种不同情形，来与已有的 LAD-ADPLASSO 方法作对比。最终不仅探究了该方法的适用条件和比较优良的预测性能，还将该方法分别用于一个医学和非医学的实际数据，发现结果与模拟情形相似，均能够得到比较优秀的预测效果。

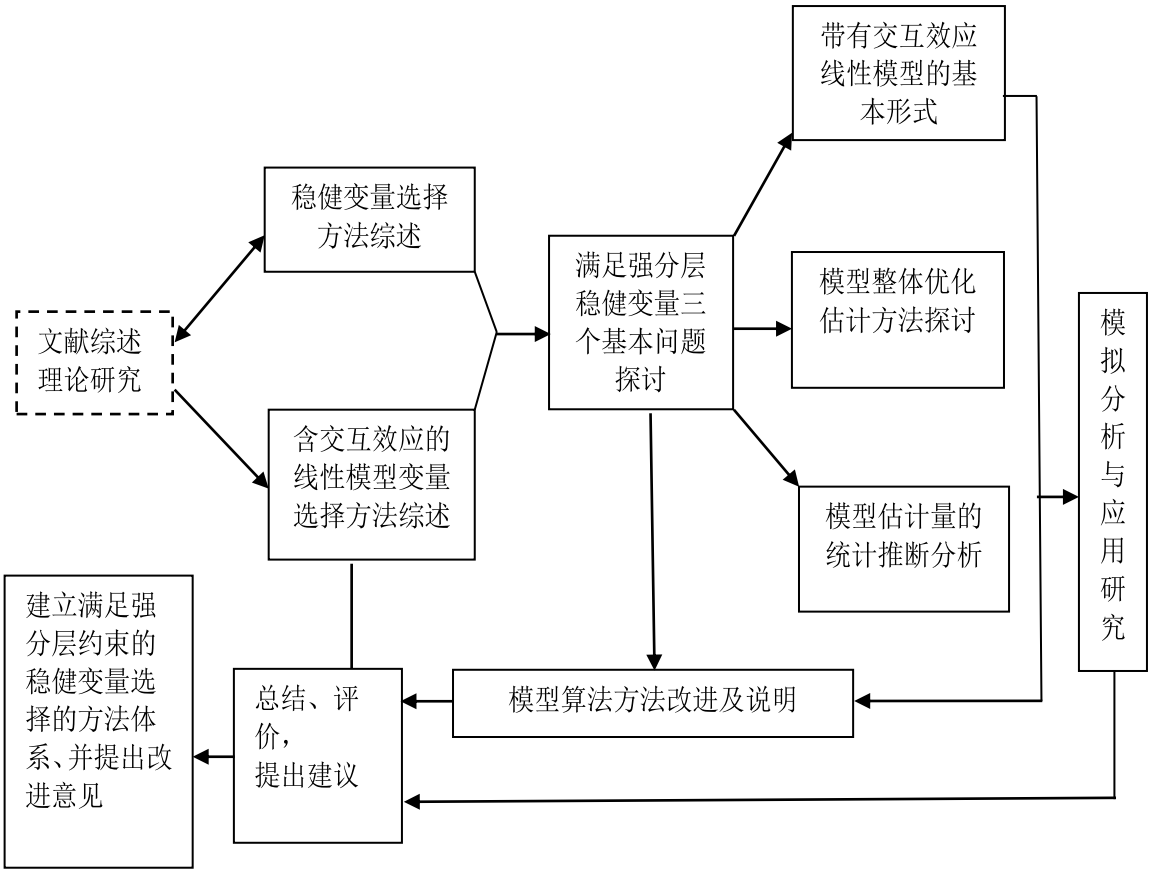

图 1.1 本文整体内容框架图

第二章 带有强分层约束的 DPD-ADPLASSO 模型形式

第一章 1.4 节较为概括的介绍了本文总体的研究内容。进一步，为了详细讨论带有交互效应的线性模型形式，及相关的解决算法改进，本章将核心内容主要分成两个子问题，即从带有交互效应的线性模型基本形式和需要的最小化目标函数两个小节来详细展开讨论。

2.1 含交互效应的线性模型形式

考虑简单线性回归模型

$$y = \sum_{j=1}^{p} x_j \beta_j + \varepsilon \tag{2.1}$$

其中$y = (y_1, \ldots, y_n)'$为响应变量（又称预测变量），$x_j = (x_{1j}, \ldots, x_{nj})'$为解释变量，$\beta = (\beta_1, \ldots, \beta_p)'$为 p 维的回归系数向量，$\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$是误差项，并且相互独立。$\varepsilon_i \sim N(0, \sigma^2)$，$j = 1, \ldots, p$，$i = 1, \ldots, n$。但在实际问题中，$x_{1,} x_{2,} \cdots x_p$并不能较为完整的体现解释变量与响应变量的相关关系。通常研究者也会关注解释变量两两之间的交互作用。例如，在诸多复杂疾病中，比如癌症，它会涉及到诸如多个基因以及环境危险因素，相关学者也会格外关注基因与基因的交互效应和基因与环境危险因素的交互效应。因此，在本文中，提出了带有主效应和交互效应的回归方程，如下所示：

$$\boldsymbol{y} = \sum_{j=1}^{p} \boldsymbol{x}_j \beta_j + \sum_{j=1}^{p-1} \sum_{j'=j+1}^{p} \boldsymbol{x}^*_{jj'} \gamma_{jj'} + \boldsymbol{\varepsilon} \tag{2.2}$$

其中$y = (y_1, \ldots, y_n)'$为响应变量，解释变量$x_j = (x_{1j}, \ldots, x_{nj})'$为非随机的确定性变量，$x^*_{jj'} = (x_{1j}x_{1j'}, \ldots, x_{nj}x_{nj'})'$为$x_j$和$x_{j'}$的交互效应，在本文是指乘积的形式，

误差项$\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$服从独立同正态分布，$\varepsilon_i \sim N(0, \sigma^2)$。不失一般性地，假定上述模型中的参数协变量$x_j$、$x^*_{jj'}$以及响应变量都是中心化的，因此模型不含截距项。以全基因组关联研究为例，模型（2.2）中的x_j通常是基因特征和部分影响形式可假定为线性的环境特征。本章研究目的为对上述模型中的参数主效应β_j、参数交互效应$\gamma_{jj'}$同时进行估计和变量选择，且保证交互效应的强分层结构成立。交互效应模型中的分层结构包括强分层和弱分层两种形式，前者要求交互效应存在的前提是相应的两个主效应都存在，而后者要求交互效应存在的前提是至少有一个相应的主效应存在即可。

Zhu(2010)在文中称模型（2.2）为强分层交互模型（SHIM）。总体来说，该模型面临如下两方面的挑战：解释性和预测准确性。通常来说，研究者希望能用尽可能少的重要的变量来解释响应变量并对之有尽可能准确的预测。解释性通常通过变量选择来实现，预测准确度则需要压缩回归估计的系数来改善，一般的，为了减小预测估计量的方差，压缩会用牺牲无偏性的办法来达到提高总体预测准确度的效果。

此外为了保证模型的估计量满足强分层结构，将采用 1.2.2 节提到的方法，将交互效应的系数写成各自对应的两个主效应的乘积形式。具体如下所示：

令交互项$\gamma_{jj'} = \alpha_{jj'}\beta_j\beta_{j'}$，即引入了中间变量$\alpha$，最终求解出的$\boldsymbol{\beta}$和$\alpha$，即相当于求解出$\boldsymbol{\beta}$和$\gamma$。即考虑下面的模型：

$$\boldsymbol{g}(\boldsymbol{\beta}) = \sum_{j=1}^{p} \boldsymbol{x}_j\beta_j + \sum_{j=1}^{p-1}\sum_{j'=j+1}^{p} \boldsymbol{x}^*_{jj'}\alpha_{jj'}\beta_j\beta_{j'} \tag{2.3}$$

则有$\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{\beta}) + \boldsymbol{\varepsilon}$。

可以发现，当改写成上述模型后，自然而然就满足了强分层约束条件。也即当某个交互效应系数不为零时，自然而然的相关的两个主效应系数也必须不能够为零。

2.2 目标函数构建及待估参数介绍

借助惩罚回归变量选择方法的思想，利用 1.2.1 节的知识，本文采用的损失函数是基于密度势差异的形式，从而达到稳健估计的效果。在回归模型中，损失函数的选择决定了参数估计的稳健性，罚函数的选择决定了变量选择的效果。从而 DPD-ADPLASSO 方法不但可以得到稳健的参数估计结果，还可以产生较好的变量选择效果。

我们最小化如下的目标函数：

$$min\frac{1}{n}\sum_{i=1}^{n}V_i+\lambda_\beta\left(|\beta_1|+\cdots+|\beta_p|\right)+\lambda_\alpha\left(|\alpha_{12}|+|\alpha_{13}|\cdots+|\alpha_{(p-1)p}|\right) \quad (2.4)$$

其中上面的V_i为每个个体的密度势差异。

$$V_i=\frac{1}{(2\pi)^{t/2}\sigma^t\sqrt{1+t}}-\frac{t+1}{t}\frac{1}{(2\pi)^{t/2}\sigma^t}e^{-t(y_i-g(\boldsymbol{x_i}))^2/2\sigma^2} \quad (2.5)$$

上式（2.4）加号的左边为第一部分，右边为第二部分。第一部分代表的是所有个体的密度势差异函数的平均值，将其作为损失函数，第二部分是普通的 Lasso 函数，λ_β和λ_α为其调节参数，右面紧随待估参数的一范数形式。

参数 t 是调节调节参数估计稳健性和有效性的一个指标。通过分析其影响函数（influence function）以及参数估计的渐进方差，t 越大越稳健，但是相对有效性越低。t越小相对有效性越高，但是稳健性越低。t取值大于 1 时，有效性相对比较低，并且当$t\to 0$时，上式（2.4）中的第一部分损失函数近似为$\frac{1}{n}\sum_{i=1}^{n}(y_i-g(\boldsymbol{x_i}))^2$，就是普通最小二乘的损失函数形式。所以综上所述，为了平衡稳健性和有效性，最终t的取值范围考虑是[0,1]。其次因为要考虑到在模拟数据和实际数据中会存在离群点比例较小的情形，那么在这种情况下，重点要关注需要得到较高的有效性，应该采用最小二乘的形式即可达到。因此在t的取值时还需要考虑在[0.001,0.1]之间的数值。最终t的取值集合为{0.001,0.01,0.03,0.05,0.07,0.09,0.1,0.2,0.4,0.6,0.8,1}。

考虑到变量选择的目的，且保证估计量具有一致性或 Oracle 性质，所以借鉴 Breiman (1995), Zou (2006), Wang, Li, and Jiang (2007) 的思想，即给不同的系数施加不同的罚权重。这种方法的核心是当估计系数比较大时，权重可以设置

小一点，则减小相应的罚强度，但当系数的估计值比较小，也就是对响应变量的影响小，则设置较大的罚权重，即相应增大罚权重。此时权重的选择有三种办法：第一种是将所有权重设为 1；第二种是借鉴 Breiman (1995) and Zou (2006) 的思想，将所有权重设为 OLS 估计下产生的系数的倒数。此时满足：

$$\omega_j^{\beta} = \frac{1}{\hat{\beta}_j^{OLS}}$$

$$\omega_{jj'}^{\gamma} = \frac{\hat{\beta}_j^{OLS}\hat{\beta}_{j'}^{OLS}}{\hat{\gamma}_{jj'}^{OLS}}$$

但是只有在当数据不存在离群点时，最小二乘估计的结果具有一致性，而当数据存在离群点时，最小二乘估计不再具有一致性。

第三种方法就是将将所有权重设为 LAD 估计下产生的系数的倒数，来替代第二种方法中使用的 OLS。因为 LAD 已经被证明具有稳健性，也即当数据中出现离群点，估计结果相比于 OLS 会准确的多，并且还具有 Oracle 的良好性质。也即此时的权重设置如下：

$$\omega_j^{\beta} = \frac{1}{\hat{\beta}_j^{LAD}}$$

$$\omega_{jj'}^{\gamma} = \frac{\hat{\beta}_j^{LAD}\hat{\beta}_{j'}^{LAD}}{\hat{\gamma}_{jj'}^{LAD}}$$

本文就采用第三种方法，即根据实际数据调节的方式来实现自适应 LASSO，也即 Adaptive Lasso 的一种变形，它已经被证明具有 Oracle 性质，借鉴 Xu(2016) 的方法，文章也是用 LAD 的估计结果作为权重。

因此最终的相应的目标函数为：

$$L = min\frac{1}{n}\sum_{i=1}^{n} V_i + \lambda_{\beta}\left(\omega_1^{\beta}|\beta_1| + \cdots + \omega_p^{\beta}|\beta_p|\right) + \lambda_{\alpha}\left(\omega_{12}^{\alpha}|\alpha_{12}| + \omega_{13}^{\alpha}|\alpha_{13}| \cdots + \omega_{(p-1)p}^{\alpha}|\alpha_{(p-1)p}|\right) \quad (2.6)$$

其中式子的V_i为：

$$V_i = \frac{1}{(2\pi)^{t/2}\sigma^t\sqrt{1+t}} - \frac{t+1}{t}\frac{1}{(2\pi)^{t/2}\sigma^t}e^{-t(y_i-g(x_i))^2/2\sigma^2} \quad (2.7)$$

总体来说，在选取合适的调整参数 t 和λ，利用（2.6）和（2.7）式需要寻找有效的算法来估计σ、β、γ的值，并且满足强分层的约束条件。

第三章 求解稳健变量选择模型的优化算法探究

3.1 结合强分层改进的稳健参数估计和变量选择算法

为了计算简便，本研究令$\lambda_\beta = \lambda_\alpha = \lambda$，于是交互效应模型的惩罚估计问题中就含有惩罚参数$\lambda$和调节参数$t$需要确定。选取最优调整参数的问题将在 3.1.4 节讨论。对于给定的调整参数λ和t，带有交互效应的线性模型，并满足强分层约束的稳健变量选择问题就可以通过如下强分层惩罚准则解决：

$$\min_{\boldsymbol{\beta},\boldsymbol{\gamma}} L(\boldsymbol{\beta},\boldsymbol{\gamma}) \quad \text{s.t. } \beta_j\beta_{j'} = 0 \Longrightarrow \gamma_{jj'} = 0, \forall j < j' \tag{3.1}$$

显然，求解上述优化问题即可得到模型待估参数 $\boldsymbol{\beta}$、$\boldsymbol{\gamma}$的估计值。

3.1.1 满足强分层约束的整体参数稳健估计算法

Xu（2016）已经将 DPD 的方法应用在线性回归模型中，采用整体迭代的算法，并证明出此算法下的估计量具有 Oracle 性质。但是由于此方法结束后得到的β、γ是不满足强分层的，本文的研究问题是对于含有交互效应的线性模型稳健的参数估计和变量选择，因此必须在算法中加入步骤，使得估计出的最终结果满足强分层的约束条件。根据 1.2 节的文献综述，借助于 Zhu(2010)提出线性模型强分层约束的（SHIM）估计方法和 Ma(2014)提出的 progressive 算法，即根据主效应的估计结果来逐步地估计交互效应，也即在求解惩罚似然问题的迭代优化算法中，只有当$\beta_j \neq 0$且$\beta_{j'} \neq 0$时才需要估计$\gamma_{jj'}$，否则直接保持$\gamma_{jj'}$等于 0 即可。固定β，采用上述估计方法和上述估计方法来估计γ，再固定γ，同样的估计方法来估计β。所以本文将此两种方法有机的结合起来，来解决强分层约

束，稳健估计和变量选择这三个子问题。

换言之，本文也将采用 Xu 文中这种整体优化估计的算法来迭代估计回归系数和方差，并且在每一步估计中使用更新系数的办法来完成强分层的约束条件。这样既保证每一步骤的稳健估计和变量选择，又满足强分层。

同时为了使得算法收敛速度较快，本文以上一个 t 值下的估计结果作为初值，这种方法也称热起点(Zeng，2015)。并且因为 t 趋近 0 时，DPD 准则即为最小二乘准则，所以此时以自适应 LASSO 的估计结果作为初值。当给定惩罚参数λ和调节参数t后，我们用迭代的算法来估计参数$\boldsymbol{\beta}$、$\boldsymbol{\gamma}$和方差σ^2。

当给定方差σ^2时，最大化

$$f_t(\boldsymbol{\beta},\boldsymbol{\gamma}|\sigma^2)$$

$$=\frac{t+1}{t}\frac{1}{(2\pi)^{t/2}\sigma^t}e^{-t(y_i-g(\boldsymbol{x_i}))^2/2\sigma^2}-n\sum_{j=1}^{p}(\lambda_{\beta j}|\beta_j|)$$

$$-n\sum_{j=1}^{p-1}\sum_{j'=j+1}^{p}(\lambda_{\gamma jj'}\gamma_{jj'}) \qquad (3.2)$$

得到β、γ的估计值。上式中只含有未知参数，但涉及非二次加权的正则化，本文借助 Tseng & Yun, 2009 的思路，采用块坐标梯度下降法求解。

给定β、γ后，最大化

$$f_t(\sigma^2|\boldsymbol{\beta},\boldsymbol{\gamma})=\frac{1}{(2\pi)^{t/2}\sigma^t\sqrt{1+t}}-\frac{t+1}{t}\frac{1}{(2\pi)^{t/2}\sigma^t}e^{-t(y_i-g(\boldsymbol{x_i}))^2/2\sigma^2} \qquad (3.3)$$

最大化上式有很多方法，比如牛顿法和二分法等。本文借鉴 Zang（2016），最终采用比较好实现的二分法来求解上式关于σ^2的光滑函数。

通过上面所说的迭代算法，在给定惩罚参数λ和调节稳健性和有效性的参数t后（通过 2.2.4 节的方法选取），迭代估计得到β、γ和方差σ^2。但是由于此方法结束后得到的β、γ是不满足强分层的，再借助于主效应的估计结果来逐步地估计交互效应，也即在求解惩罚似然问题的迭代优化算法中，只有当$\beta_j\neq 0$且$\beta_{j'}\neq 0$时才需要估计$\gamma_{jj'}$，否则直接保持$\gamma_{jj'}$等于 0 即可。固定β，采用上述估计方法和上述估计方法来估计γ，再固定γ，同样的估计方法来估计β，直至目标函数收敛。从而模型在整体上估计参数的算法见 3.1 节最后的算法 3.1。

3.1.2 回归系数的 BCGD 估计算法

$\boldsymbol{\beta},\boldsymbol{\gamma}$的估计是用块坐标梯度下降算法(Tseng & Yun, 2009)。对于 DPD-ADPLASSO 的损失函数，由于它不是一个正定的矩阵。采用普通的二阶泰勒展开方法导致每一次迭代进行求解时，泰勒展开后不一定存在最优解。基于此，本文采用块坐标梯度下降算法（block coordinate gradient descent, BCGD）。块坐标梯度下降算法主要用来解决损失函数可导并且罚函数可分或者分块可分的优化问题。损失函数可以是凸函数或者凹函数也可以是非凸函数或者非凹函数，同时为了保证块坐标梯度下降算法的全局收敛性，每一次更新的坐标需要采用一定的准则进行选取。BCGD 算法的核心思想是先将目标函数进行二阶泰勒展开，并且其二阶倒数用一个正定或者负定的矩阵替代，进而求得下降方向，再按照相应的准则求出最优步长，然后就可以得到下一步的初始估计值。并且参照 Tseng & Yun 的文章思路，为了使得保证该算法的具有线性的收敛速度，本文 BCGD 估计算法整体如下：首先，先采用 *Gauss-Southwell-*γ 准则选定需要更新的坐标集合 *J*, 选取负定矩阵 *H*，接着，利用坐标下降算法求得下降方向 *d*，然后，最优步长 *s* 采用 *Arimijo* 进行选取，最后，结合$\hat{\beta}^t = \hat{\beta}^t + sd$则得到下一步迭代时的初值。

3.1.3 方差的估计算法

方差 σ 估计可以用牛顿迭代法或者二分法，由于在固定 $\boldsymbol{\beta},\boldsymbol{\gamma}$ 后，目标函数只有一个未知数，也即是光滑的目标函数。牛顿迭代法的核心思路是将非线性方程通过泰勒展开线性化，并且具有二阶收敛速度，但是该方法有明显的缺点，除了要求目标函数可导外，还对初值会很敏感，因此在初值方面要求苛刻（张晓勇，2013），相比于牛顿迭代法，用对初值相对不敏感的二分法来估计方差 σ会更加合适。同时该方法在估计方差 σ也有运用（见 Zang，2016）。因此，综上所示，本文最终采用二分法来估计方差σ。并且在编写 R 程序实现时，二分法也比牛顿迭代法方便快捷。

3.1.4 调整参数 λ 和 t 的估计算法

（1）调整参数 t 的选取，根据最小化模型误差得到

由于在给定一个 t 值后，便可根据算法 3.1 中迭代的算法估计出主效应系数和交互效应系数。在模拟时，我们根据模型误差最小为准则，选取使得整体的模型误差最小时所对应的 t 值，在实际数据的应用时，由于实际模型未知，借助于 Xu,2016 的方法，采用最大化方差的有效性准则，挑选最优的 t 值。

（2）根据BIC准则选取惩罚参数 λ

惩罚参数存在很多种方法进行选取。例如，交叉验证，广义交叉验证，AIC 准则以及 BIC 准则。为了使得计算简便，并且变量选择结果具有一致性，我们这里采用最小化 BIC 的一个准则来选取。根据 BIC 的思路，最小化：

$$\sum_{i=1}^{n}\frac{1}{(2\pi)^{t/2}\sigma^{t}\sqrt{1+t}}-\frac{t+1}{t}\frac{1}{(2\pi)^{t/2}\sigma^{t}}e^{-t(y_i-g(x_i))^2/2\sigma^2}+n\sum_{j=1}^{p}(\lambda_{\beta j}|\beta_j|)+n\sum_{j=1}^{p-1}\sum_{j'=j+1}^{p}(\lambda_{\gamma jj'}\gamma_{jj'})-\sum_{j=1}^{p-1}\sum_{j'=j+1}^{p}\sum_{j=1}^{p-1}\sum_{j'=j+1}^{p}(\lambda_{\gamma jj'}\gamma_{jj'}) \tag{3.4}$$

为了使得 DPD-ADPLASSO 稳健变量选择的估计结果具有 Oracle 性质，本文以 LAD 估计结果替代，即最终的计算公式为：

$$\lambda_{\beta j}=\frac{log(n)}{n|\beta_j^{LAD}|} \tag{3.5}$$

$$\lambda_{\gamma jj'}=\frac{log(n)}{n|\gamma_{jj'}^{LAD}|} \tag{3.6}$$

其中，β_j^{LAD}和$\gamma_{jj'}^{LAD}$是 LAD 的估计结果。

综合 3.1.1 节到 3.1.4 节，本文提出的稳健估计的优化算法可以用算法 3.1 中的伪代码来表示。

算法 3.1：求解算法

输 入：协变量$x_j, x_{jj'}^* \in \mathbb{R}^n$，$j < j', j, j' = 1, \ldots, p$；响应变量$y \in \mathbb{R}^n$；调节参数 t 的取值是$t\epsilon\{0.001,0.01,0.03,0.05,0.07,0.09,0.1,0.2,0.4,0.6,0.8,1\}$。

初始化：令$\hat{\beta}^{(0)}$、$\hat{\gamma}^{(0)}$分别为 LAD 法的参数初始估计值，进而利用公式（3.5）和（3.6）计算λ_{nj}。令迭代指标$m = 1$。

重复如下步骤(1)-(3)直到目标函数$L(\beta, \gamma)$收敛：对每一个$i\epsilon\{1,2,\cdots s\}$, s 为 t 取值集合元素的总个数。

（1）取 t 集合的第 *i* 个作为 t 的取值

（2）迭代估计 $\hat{\beta}^{(m)}$, $\hat{\gamma}^{(m)}$和$\sigma^{(m)}$

（3）求解系数估计值，结合 SHIM 和 progressive 估计方法，分别固定β更新γ，固定γ更新β。计算每一个 t 值所对应的模型误差，选取使得误差最小的对应 t 值

令$m = m + 1$，分别计算$L\left(\hat{\beta}^{(m)}, \hat{\gamma}^{(m)}\right)$和$L\left(\hat{\beta}^{(m-1)}, \hat{\gamma}^{(m-1)}\right)$

收敛准则为：

$$\frac{\left|L\left(\hat{\beta}^{(m)}, \hat{\gamma}^{(m)}\right) - L\left(\hat{\beta}^{(m-1)}, \hat{\gamma}^{(m-1)}\right)\right|}{\left|L\left(\hat{\beta}^{(m-1)}, \hat{\gamma}^{(m-1)}\right)\right|} < 10^{-4}$$

3.2 模型的统计推断

在实际数据应用本文提出的此方法进行参数估计和变量选择时，会不可避免的出现错误估计。比如真实重要的解释变量被错误的剔除，也有可能真实不重要的解释变量被错误的选出。这些情况都会对模型的准确估计带来挑战。为了进一步的评价模型的估计效果，本节将采用 Bootstrap 重抽样方法来做统计推断，即在 100 次 Bootstrap 重抽样后构造出回归系数β_j和$\gamma_{jj'}$的置信区间。

假定解释变量$\boldsymbol{x}_j$是确定性非随机的，根据 Freedman (1981)，此时对于线性回归模型标准的 Bootstrap 抽样过程为残差 Bootstrap。对初始模型，假定求解优化问题得到的参数惩罚估计为$\widehat{\boldsymbol{\beta}}$、$\widehat{\boldsymbol{\gamma}}$，于是模型的残差是：

$$\boldsymbol{e} = \boldsymbol{y} - \sum_j \boldsymbol{x}_j \hat{\beta}_j - \sum_{j<j'} \boldsymbol{x}^*_{jj'} \hat{\gamma}_{jj'} \tag{3.7}$$

然后中心化残差$\{e_i - \bar{e}_n, i = 1, \dots, n\}$，其中$\bar{e}_n = n^{-1}\sum_{i=1}^{n} e_i$。将中心化后的残差进行有放回的随机抽样，得到 Bootstrap 残差$e^\dagger = \left(e_1^\dagger, \dots, e_n^\dagger\right)'$。于是 Bootstrap 样本新的响应变量为

$$\boldsymbol{y}^\dagger = \sum_j \boldsymbol{x}_j \hat{\beta}_j + \sum_{j<j'} \boldsymbol{x}^*_{jj'} \hat{\gamma}_{jj'} + \boldsymbol{e}^\dagger \tag{3.8}$$

然后对 Bootstrap 数据集$\left\{y^\dagger, x_j, x^*_{jj'}: j < j', j, j' = 1, \dots, p\right\}$建立强分层惩罚问题并求得参数$\beta$、$\gamma$的 Bootstrap 惩罚估计$\beta^\dagger$、$\gamma^\dagger$。

假设重复上述 Bootstrap 的有放回随机中心化残差抽样过程C次，得到的参数惩罚估计分别为$\beta_c^\dagger$、$\gamma_c^\dagger$，$c = 1, \dots, C$。按照 Chatterjee & Lahiri (2011)，令$T_{c,j} = \sqrt{n}\left(\beta_{c,j}^\dagger - \hat{\beta}_j\right)$，$\hat{t}_j(\alpha)$为$\left\{\left|T_{c,j}\right|, c = 1, \dots, C\right\}$的$1-\alpha$百分位数，则可以构造回归系数$\beta_j$的置信度为$1-\alpha$的 Bootstrap 置信区间为

$$I_\alpha(\beta_j) = \left\{t \in \mathbb{R}: \left|t - \hat{\beta}_j\right| \le n^{-1/2}\hat{t}_j(\alpha)\right\} \tag{3.9}$$

同理可构建$\gamma_{jj'}$的置信度为$1-\alpha$的置信区间$I_\alpha(\gamma_{jj'})$。

第四章 验证算法有效性的数值模拟

本章进行模拟研究。主要是为了评价 DPD-ADPLASSO 稳健变量选择方法在带有交互效应的普通线性模型上满足强分层约束条件，在不同离群点类型(type)，离群点比例(rate)以及不同的样本量(n)下变量选择效果以及预测效果。另外，模拟研究中还考虑其他两种估计方法：一种是基于 LAD-ADPLASSO 的稳健变量选择方法，另一种是以 MM 估计的普通变量选择方法，也称为 Oracle 方法，Oracle 方法是将真正起作用的解释变量纳入回归方程，从而进行稳健估计和预测。具体参数设置如下：

4.1 模拟参数设置

本节离群点比例取值集合为(5%；10%；20%)。设置不同离群点比例，主要是为了比较 DPD-ADPLASSO 模型和另外两种模型的结果。同时为了考察，在样本量较小，适中和较大情形下，DPD-ADPLASSO 模型的表现，样本量取值集合为(200；400；800)。每种样本量情形下都假定真实模型满足强分层结构，在协变量取值集合上主要参考 Zeng (2015)，即 p=10 个参数主效应协变量和 $p(p-1)/2=45$ 个参数的交互效应协变量，其中在十个主效应协变量里前六个主效应是显著的，显著影响因变量，交互效应里有六个对因变量起重要作用的变量。所有主效应和非零交互效应系数具体取值分别如下：

$$(\beta_1,\beta_2,\beta_3,\beta_4,\beta_5,\beta_6,\beta_7,\beta_8,\beta_9,\beta_{10})=(5,4,3,2,1,0.5,0,0,0,0)$$

$$(\gamma_{12},\gamma_{13},\gamma_{23},\gamma_{24},\gamma_{34},\gamma_{56})=(3,2.5,2.0,1.5,1.0,0.5)$$

调节参数 *t* 取值集合为 $(0.001,0.01,0.03,0.05,0.07,0.09,0.1,0.2,0.4,0.6,0.8,1.0)$，如算法 3.1 所示。该取值集合主要参考 Basu et al., (1998)。取值集合中包含一个很小的值，主要是因为样本中离群点不是特别明显的情况时，损失函数即近似为普通的 OLS，因此 *t* 取了一个极小值 0.001，从而保证 DPD-ADPLASSO 同样可以获得较好的效果。

最后，五种不同离群点的情形设置如下:

情形一：离群点来自响应变量。非离群点解释变量服从$N(0,I_n)$，非离群点误差项也服从标准正态分布，离群点误差项服从 *laplace*（0,10）分布。

情形二：离群点来自解释变量。离群点解释变量$\boldsymbol{x}=\boldsymbol{x}+(e_1,e_2,\cdots e_{10})$,其中$e_5\sim N(0,5)$; $e_i\sim N(0,1), i=1,2,3,4,6,7,8,9,10$,误差项服从标准正态分布。非离群点解释变量见情形一。此种离群点的设置情况主要参考 Leng(2010)。

情形三：离群点同时是来自解释变量和响应变量。非离群点解释变量同情形一，非离群点误差项服从标准正态。来自解释变量的离群点的生成机制同情形一的生成机制，来自响应变量的离群点的生成机制同情形二的生成机制。

情形四：离群点同时是来自解释变量和响应变量。非离群点解释变量带状结构，即$x_i\sim N(0，\Sigma)$, Σ中的（i，j）元素是$0.4^{|i-j|}$，当且仅当$|i-j|=1$，否则为 0，非离群点误差项服从标准正态。来自解释变量的离群点的生成机制同情形一的生成机制，来自响应变量的离群点的生成机制同情形二的生成机制。

情形五：离群点同时是来自解释变量和响应变量。解释变量x_i服从均值为 0，方差为 0.4 指数衰减的正态分布，即$x_i\sim N(0，\Sigma)$, Σ中的（i，j）元素是$0.4^{|i-j|}$。来自解释变量的离群点的生成机制同情形一的生成机制，来自响应变量的离群点的生成机制同情形二的生成机制。

对每种情形，模拟研究重复进行 100 次。

4.2 变量选择和参数估计效果

为评价模型变量选择效果，本研究考虑所拟合的模型中回归系数非零的协变量平均被选择出来的个数所占的比例（True Positive Rate, TPR）和非零协变量被正确剔除的个数所占的比例（True Negative Rate, TNR）。例如，参数主效应部分的指标计算分别为

$$TP(\hat{\boldsymbol{\beta}})=\sum_j I(\hat{\beta}_j\neq 0)*I(\beta_j\neq 0)$$

和

$$TN(\hat{\boldsymbol{\beta}})=\sum_j I(\hat{\beta}_j= 0)*I(\beta_j= 0)$$

参数交互效应的指标TP($\hat{\gamma}$)、TN($\hat{\gamma}$)的定义类似。所以此值越接近于 1 说明变量选择效果越准确。TPR 也称灵敏度，TNR 也称特异度。但在医学数据中，也会关注分类的准确度（accuracy, ACC），分类的准确度是 TP 与 TN 的综合衡量效果，准确度的计算如下:

$$ACC = \frac{TP + TN}{P_{all}}$$

此值越接近于 1 说明总体变量选择的准确度越高。

预测效果分为两方面，参数β和γ的估计效果用均方误差（Mean Squares Error，MSE）来评价，其定义为

$$MSE(\hat{\boldsymbol{\beta}}) = \sum_{j=1}^{p}(\hat{\beta}_j - \beta_j)^2 = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2$$

$MSE(\hat{\boldsymbol{\gamma}})$类似定义。下面 4.3 节表中的 MSE 为β和γ的 MSE 之和。

此外，模型整体误差 (Fan & Li, 2001)中位数(Median，MED)也可来衡量预测效果，MSE 和 MED 的值越小，说明模型的预测效果越好。其中，模型误差定义为：

$$ME(\beta_n) = (\beta_n - \beta_o)^T E(xx^T)(\beta_n - \beta_0)$$

ME(γ_n)类似定义。下面 4.3 节表中的 MED 也是为β和γ的 MED 之和。

4.3 模拟结果分析

分析模拟结果可以有如下结论：

一、DPD- ADPLASSO 稳健的变量选择自身的效果

固定离群点比例，分别观察五种不同的离群点情形，发现在这五种离群点情形下，随着样本量的增大，主效应和交互效应的 TP 和 TN 的值越来越接近于 1，选择的总体准确度 ACC 值也越来越接近与 1，说明变量选择的效果具有一致性；此外 MSE 和 MED 也呈现越来越小的趋势，与 Oracle 方法表现的 MSE 和MED越来越接近。这体现了DPD-ADPLASSO稳健变量选择估计量的一致性。结合变量选择效果和模型预测的效果，说明 DPD-ADPLASSO 可以得到一致性

的结果。

二、DPD-ADPLASSO 与 LAD-ADPLASSO 稳健变量选择结果的比较

（1）通过固定离群点情形和样本量，不难发现，在每一种情形下，发现无论离群比例怎样变化，DPD-ADPLASSO 的 MSE 和 MED 都比 LAD-ADPLASSO 小，也即在综合的模型预测效果上有非常好的表现。

（2）对于第四种和第五种离群点情形：当样本量较小时，就 TPR 而言，DPD-ADPLASSO 差于 LAD-ADPLASSO，因为 DPD-ADPLASSO 牵涉到方差的一致估计 (方差的估计不一定收敛到真值)，所以样本量较小时，TPR 差于 LAD-ADPLASSO 方法可以理解，但是其它方面，如 TNR，ACC 都要比 LAD-ADPLASSO 要好。然后，随着样本量增大，DPD-ADPLASSO 的 TPR 值越来越接近于 1；在特异度方面，从各个表可以看出，DPD-ADPLASSO 要优于 LAD-ADPLASSO。

三、离群比例影响 DPD- ADPLASSO 与 LAD-ADPLASSO 的表现效果

从表 4.1-4.5 可以看出明显的一点：固定样本量和离群点类型，随着离群比例的增大，DPD-ADPLASSO 和 LAD-ADPLASSO 的 MSE 和 MED 也呈现越来越大的趋势，并且 TP 和 TN 也逐渐缩小（特别是表 4.5 非常明显），而且，相比于 DPD-ADPLASSO 的 TN 值，LAD-ADPLASSO 则下降相对更快。但是随着样本量的增大，可以看到这种下降的幅度越来越小。

此外，观察解释变量的相关结构对 DPD-ADPLASSO 与 LAD-ADPLASSO 的影响。从表 4.3-4.5 中可以看出，固定样本量和离群比例，随着相关结构从独立到慢慢相关性增强，MSE 和 MED 也逐渐增大，但总体的选择准确度保持在 0.990 的水平上。

表 4.1　第一种离群点情形下参数估计与模型预测结果

n	rate	方法	TPR	TNR	ACC	MSE	MED
			P=10				
200	0.05	DPD	1.000	0.997	0.998	0.081	0.436
		LAD	1.000	0.986	0.989	0.085	0.516
		Oracle	1.000	1.000	1.000	0.074	0.062
	0.1	DPD	0.999	0.997	0.997	0.094	0.467
		LAD	1.000	0.972	0.978	0.099	0.513
		Oracle	1.000	1.000	1.000	0.086	0.075
	0.2	DPD	0.984	0.996	0.994	0.127	0.588
		LAD	1.000	0.952	0.963	0.134	0.580
		Oracle	1.000	1.000	1.000	0.117	0.094
400	0.05	DPD	0.999	0.999	0.999	0.059	0.153
		LAD	1.000	0.992	0.994	0.064	0.193
		Oracle	1.000	1.000	1.000	0.035	0.032
	0.1	DPD	1.000	0.999	0.999	0.063	0.201
		LAD	1.000	0.990	0.993	0.071	0.247
		Oracle	1.000	1.000	1.000	0.038	0.034
	0.2	DPD	1.000	1.000	1.000	0.072	0.201
		LAD	1.000	0.981	0.985	0.085	0.246
		Oracle	1.000	1.000	1.000	0.047	0.043
800	0.05	DPD	1.000	1.000	1.000	0.037	0.087
		LAD	1.000	0.997	0.997	0.049	0.106
		Oracle	1.000	1.000	1.000	0.018	0.018
	0.1	DPD	1.000	1.000	1.000	0.046	0.100
		LAD	1.000	0.998	0.998	0.057	0.114
		Oracle	1.000	1.000	1.000	0.019	0.017
	0.2	DPD	1.000	1.000	1.000	0.058	0.108
		LAD	1.000	0.994	0.996	0.085	0.134
		Oracle	1.000	1.000	1.000	0.024	0.021

表 4.2　第二种离群点情形参数估计与模型预测结果

n	rate	方法	TPR	TNR	ACC	MSE	MED
			P=10				
200	0.05	DPD	1.000	0.996	0.997	0.065	0.486
		LAD	1.000	0.987	0.989	0.076	0.529
		Oracle	1.000	1.000	1.000	0.063	0.057
	0.1	DPD	1.000	0.994	0.996	0.068	1.010
		LAD	1.000	0.987	0.990	0.086	1.232
		Oracle	1.000	1.000	1.000	0.057	0.055
	0.2	DPD	0.998	0.997	0.997	0.182	3.103
		LAD	1.000	0.987	0.990	0.219	3.510
		Oracle	1.000	1.000	1.000	0.053	0.062
400	0.05	DPD	1.000	0.998	0.999	0.034	0.348
		LAD	1.000	0.994	0.996	0.038	0.393
		Oracle	1.000	1.000	1.000	0.033	0.033
	0.1	DPD	1.000	0.999	0.999	0.061	0.928
		LAD	1.000	0.995	0.996	0.075	1.007
		Oracle	1.000	1.000	1.000	0.026	0.029
	0.2	DPD	1.000	1.000	1.000	0.196	3.161
		LAD	1.000	0.995	0.996	0.229	3.452
		Oracle	1.000	1.000	1.000	0.022	0.028
800	0.05	DPD	1.000	1.000	1.000	0.021	0.247
		LAD	1.000	0.998	0.999	0.025	0.270
		Oracle	1.000	1.000	1.000	0.015	0.015
	0.1	DPD	1.000	1.000	1.000	0.064	0.762
		LAD	1.000	0.998	0.998	0.075	0.825
		Oracle	1.000	1.000	1.000	0.013	0.015
	0.2	DPD	1.000	1.000	1.000	0.226	2.929
		LAD	1.000	0.999	0.999	0.246	3.050
		Oracle	1.000	1.000	1.000	0.012	0.015

表 4.3　第三种离群点情形参数估计与模型预测结果

n	rate	方法	TPR	TNR	ACC	MSE	MED
			P=10				
200	0.05	DPD	1.000	0.987	0.990	0.156	0.609
		LAD	1.000	0.947	0.959	0.185	0.895
		Oracle	1.000	1.000	1.000	0.104	0.087
	0.1	DPD	0.984	0.981	0.982	0.131	1.534
		LAD	1.000	0.886	0.911	0.210	2.207
		Oracle	1.000	1.000	1.000	0.127	0.112
	0.2	DPD	0.912	0.975	0.961	0.357	4.563
		LAD	1.000	0.792	0.837	0.625	8.117
		Oracle	1.000	1.000	1.000	0.202	0.241
400	0.05	DPD	0.999	0.986	0.989	0.049	0.601
		LAD	1.000	0.972	0.978	0.050	0.769
		Oracle	1.000	1.000	1.000	0.047	0.043
	0.1	DPD	1.000	0.984	0.988	0.095	1.587
		LAD	1.000	0.941	0.954	0.132	2.148
		Oracle	1.000	1.000	1.000	0.054	0.052
	0.2	DPD	0.975	0.976	0.976	0.204	4.689
		LAD	1.000	0.867	0.896	0.373	6.579
		Oracle	1.000	1.000	1.000	0.093	0.112
800	0.05	DPD	1.000	0.989	0.991	0.029	0.565
		LAD	1.000	0.982	0.986	0.043	0.645
		Oracle	1.000	1.000	1.000	0.022	0.019
	0.1	DPD	1.000	0.980	0.985	0.086	1.675
		LAD	1.000	0.969	0.976	0.115	1.932
		Oracle	1.000	1.000	1.000	0.029	0.027
	0.2	DPD	0.979	0.961	0.965	0.258	4.630
		LAD	1.000	0.923	0.940	0.314	5.819
		Oracle	1.000	1.000	1.000	0.044	0.054

表 4.4　第四种离群点情形参数估计与模型预测结果

n	rate	方法	TPR	TNR	ACC	MSE	MED
			P=10				
200	0.05	DPD	0.996	0.978	0.982	0.106	0.831
		LAD	1.000	0.923	0.939	0.119	1.125
		Oracle	1.000	1.000	1.000	0.104	0.077
	0.1	DPD	0.988	0.977	0.979	0.138	1.489
		LAD	1.000	0.875	0.902	0.265	2.367
		Oracle	1.000	1.000	1.000	0.136	0.119
	0.2	DPD	0.928	0.975	0.965	0.284	4.141
		LAD	1.000	0.775	0.824	0.611	8.889
		Oracle	1.000	1.000	1.000	0.210	0.241
400	0.05	DPD	1.000	0.977	0.982	0.055	0.831
		LAD	1.000	0.967	0.975	0.075	1.018
		Oracle	1.000	1.000	1.000	0.046	0.038
	0.1	DPD	0.999	0.989	0.991	0.094	1.471
		LAD	1.000	0.930	0.945	0.132	2.143
		Oracle	1.000	1.000	1.000	0.056	0.057
	0.2	DPD	0.968	0.970	0.970	0.217	3.981
		LAD	1.000	0.857	0.888	0.376	6.218
		Oracle	1.000	1.000	1.000	0.096	0.108
800	0.05	DPD	1.000	0.988	0.991	0.034	0.758
		LAD	1.000	0.983	0.987	0.044	0.885
		Oracle	1.000	1.000	1.000	0.026	0.021
	0.1	DPD	1.000	0.976	0.981	0.093	1.273
		LAD	1.000	0.961	0.970	0.122	1.536
		Oracle	1.000	1.000	1.000	0.028	0.027
	0.2	DPD	0.993	0.964	0.970	0.226	4.097
		LAD	1.000	0.910	0.930	0.319	5.362
		Oracle	1.000	1.000	1.000	0.048	0.055

表 4.5　第五种离群点情形参数估计与模型预测结果

n	rate	方法	TPR	TNR	ACC	MSE	MED
			P=10				
200	0.05	DPD	0.994	0.985	0.987	0.107	0.763
		LAD	1.000	0.938	0.952	0.108	0.986
		Oracle	1.000	1.000	1.000	0.104	0.078
	0.1	DPD	0.979	0.978	0.978	0.129	1.893
		LAD	1.000	0.893	0.917	0.169	2.599
		Oracle	1.000	1.000	1.000	0.118	0.101
	0.2	DPD	0.908	0.978	0.963	0.308	4.831
		LAD	1.000	0.796	0.840	0.593	8.085
		Oracle	1.000	1.000	1.000	0.185	0.216
400	0.05	DPD	1.000	0.989	0.991	0.054	0.662
		LAD	1.000	0.970	0.977	0.057	0.814
		Oracle	1.000	1.000	1.000	0.049	0.042
	0.1	DPD	0.998	0.981	0.985	0.091	1.472
		LAD	1.000	0.934	0.948	0.142	2.108
		Oracle	1.000	1.000	1.000	0.059	0.064
	0.2	DPD	0.968	0.976	0.974	0.244	4.359
		LAD	1.000	0.879	0.905	0.382	6.453
		Oracle	1.000	1.000	1.000	0.095	0.119
800	0.05	DPD	1.000	0.987	0.990	0.042	0.856
		LAD	1.000	0.988	0.991	0.054	0.980
		Oracle	1.000	1.000	1.000	0.024	0.021
	0.1	DPD	0.998	0.972	0.977	0.106	1.810
		LAD	1.000	0.971	0.977	0.122	2.007
		Oracle	1.000	1.000	1.000	0.028	0.027
	0.2	DPD	0.987	0.964	0.969	0.202	4.890
		LAD	1.000	0.918	0.936	0.275	5.986
		Oracle	1.000	1.000	1.000	0.045	0.059

进一步，为了从直观上展示 DPD-ADPLASSO 和 LAD-ADPLASSO 在 MSE 与 MED 上的差异，本文选取当样本量为 800 时，分别展示不同的下两种方法的效果差异。图 4.1 和图 4.2 依次离群比例为 0.05, 0.2 时两种方法的估计准确度。横轴数字 1-5 表示的是 4.1 节的五种不同离群情形，纵轴为各自指标的数值结果。

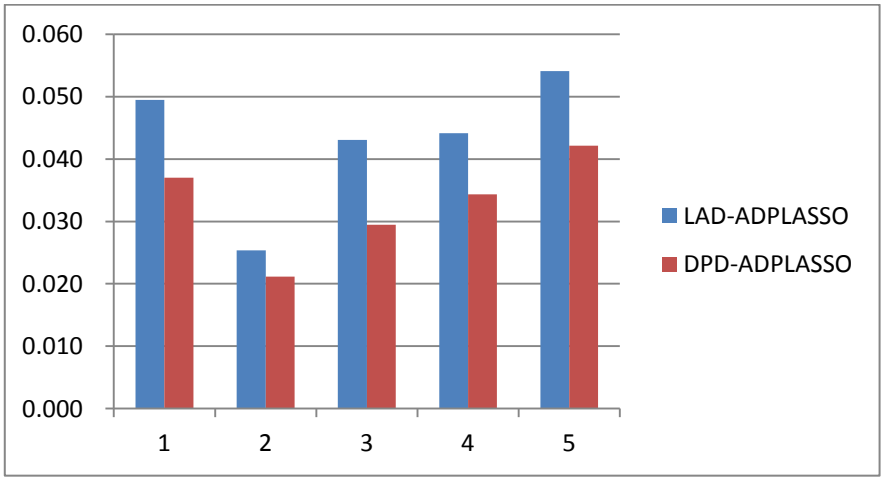

图 4.1 在离群比例为 0.05 时不同离群情形下 MSE 的差异

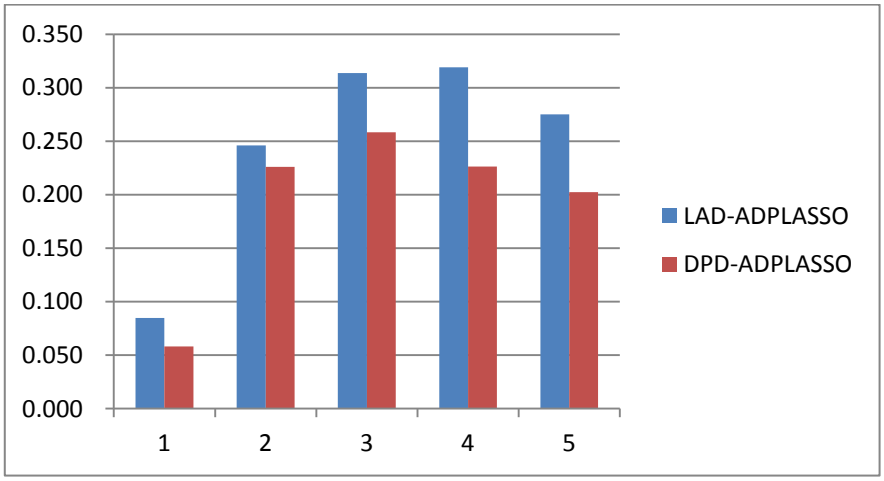

图 4.2 在离群比例为 0.2 时不同离群情形下 MSE 的差异

从上述图 4.1 和 4.2 可以看出，不论何种离群点情形，DPD-ADPLASSO 的估计准确度都要比 LAD-ADPLASSO 要好。并且上下两幅图的纵轴尺度增大，这也表明，当离群比例减小时，估计准确度会越来越好。

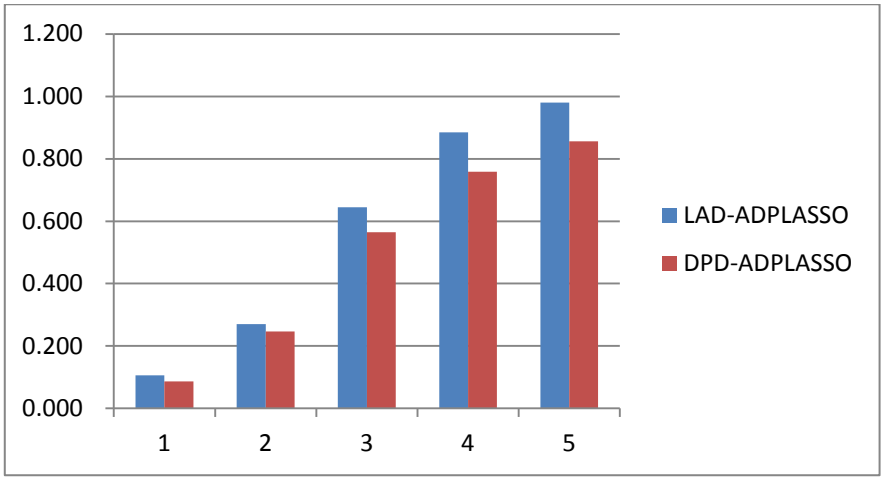

图 4.3 在离群比例为 0.05 时不同离群情形下 MED 的差异

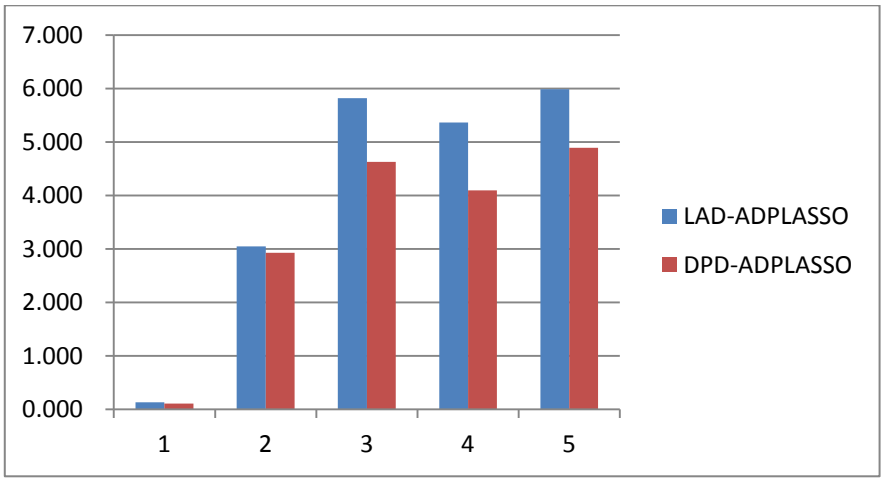

图 4.4 在离群比例为 0.2 时不同离群情形下 MED 的差异

下图 4.3 和 4.4 是从预测精度的角度来展示 DPD-ADPLASSO 和 LAD-ADPLASSO 在 MSE 与 MED 上的差异。从这两幅图可以看出随着离群比例的增大，两种方法的 MED 都呈现增大的趋势，此外，不论是何种离群比例设置情形，DPD-ADPLASSO 的模型误差精确度都要比 LAD-ADPLASSO 要好，此外，也可以看出当解释变量和响应变量均有离群点时，MED 值都是要显著高于解释变量或响应变量单独有离群点时 MED 的值。

此外，到目前为止，我们一直限定数据是满足强分层的假设前提，但是在实际数据中，研究者的第一步是需要先判断数据是否满足强分层，因此，为了比较数据在不满足强分层结构和满足强分层结构下的变量选择精度和预测准确度，本文增加考虑一种模拟设置，即样本量取值 1600，这也和 5.2 节的实际数据的样本量保持一致，主效应和交互效应的取值与本节伊始的设置相同。因此综合比较 DPD-ADPLASSO 在强分层和不分层结构里在五种不同离群情形下的估计效果。重复模拟 100 次，估计结果见表 4.6。

表 4.6　n=1600, p=10,rate=0.10 的模拟结果

情形	类型	TPR	TNR	ACC	MSE	MED
情形一	强分层	1.000	1.000	1.000	0.004	0.045
	不分层	1.000	1.000	1.000	0.003	0.046
情形二	强分层	1.000	1.000	1.000	0.059	0.767
	不分层	1.000	1.000	1.000	0.059	0.777
情形三	强分层	1.000	0.984	0.988	0.053	0.660
	不分层	1.000	0.986	0.989	0.061	0.694
情形四	强分层	1.000	0.976	0.981	0.080	1.514
	不分层	1.000	0.981	0.985	0.088	1.524
情形五	强分层	1.000	0.970	0.977	0.091	1.801
	不分层	0.999	0.973	0.978	0.102	1.824

从表 4.6 可以看出，当离群值单独来自于响应变量或者单独来自于解释变量时，TPR,TNR 以及 ACC 保持在较高的水平，并且 MSE 在强分层和不分层的结果下差别不大，但是预测精度 MED 强分层要优于不分层的结果。当离群值同时来源于解释变量和响应变量时，随着解释变量相关结构复杂度的增加，TPR,TNR 以及 ACC 逐步减小，MSE 和 MED 呈现逐渐增加的趋势。具体到此种情况的每种情形，发现 TNR 在强分层上稍微低于不分层的结果，但从 MSE 和 MED 上来看，明显强分层要优于不分层结构下的结果，总体来说，在五种情形下，强分层的变量选择准确度和不分层的变量选择准确度相差不大，但从预测角度上观察均方误差和模型的整体误差，强分层结果要优于不分层的结果。

4.4 参数统计推断

本节对于$p=10$，且样本量$n=400$，离群点类型见 4.1 节的情形五，即离群点是来自解释变量和响应变量，且解释变量存在相关结构，服从均值为 0，方差为 0.4 指数衰减的正态分布，生成一个模拟数据集，主效应和交互效应系数具体取值如下：

$$(\beta_1,\beta_2,\beta_3,\beta_4,\beta_5,\beta_6,\beta_7,\beta_8,\beta_9,\beta_{10})=(5,4,3,2,1,0.5,0,0,0,0)$$

$$(\gamma_{12},\gamma_{13},\gamma_{23},\gamma_{24},\gamma_{34},\gamma_{56})=(3,2.5,2.0,1.5,1.0,0.5)$$

接着采用 100 次 Bootstrap 重抽样方法来构造β_j和$\gamma_{jj'}$的置信区间，β_j和$\gamma_{jj'}$的置信区间如下图 4.5，其中下图 4.5 中只画出参数点估计值非零的协变量。

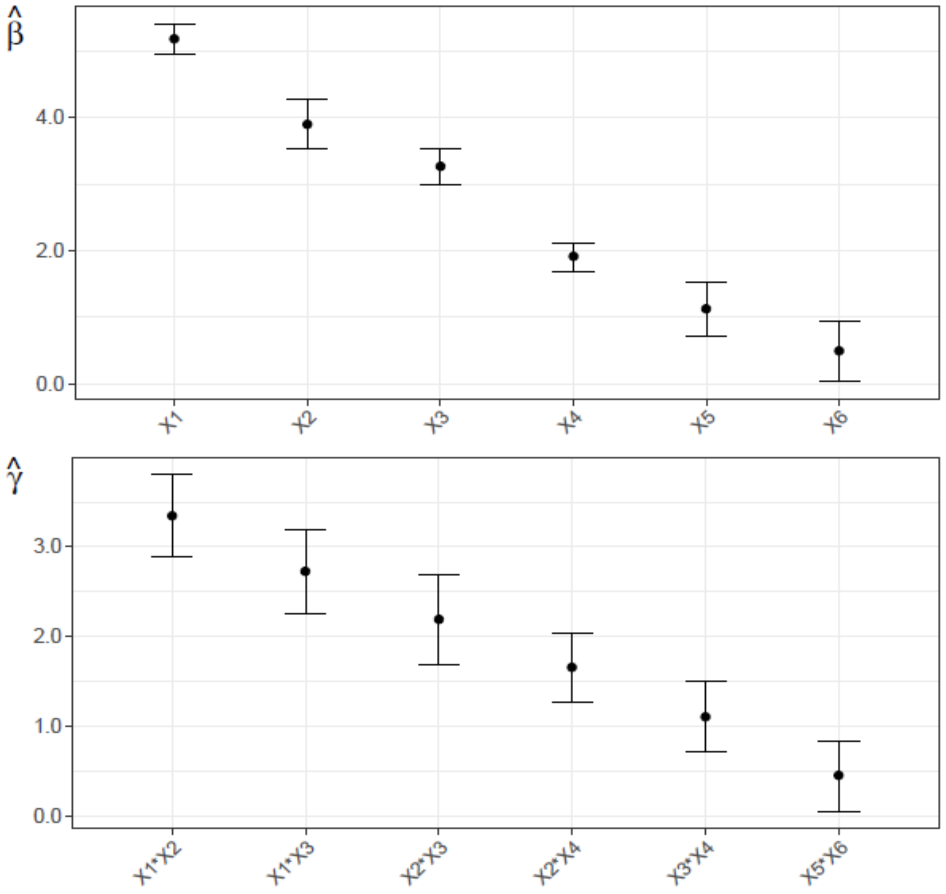

图 4.5　离群比例为 0.1 时基于 Bootstrap 构建的β_j和$\gamma_{jj'}$的置信区间

上述置信区间图 4.5 的上半部分为主效应系数估计值非零的部分，横轴是主效应的解释变量名，主效应系数的真实值是只有前六个变量的系数非零，依次为 5,4,3,2,1,0.5。从上图可以看出第一个到第六个自变量的估计值的置信区间确实包含了真实的主效应系数值，当然也准确选择出了非零的主效应；图 4.5 的下半部分是交互效应系数的估计值，横轴为交互效应解释变量，纵轴为交互效应系数估计值，不难可以看出不仅正确选出了非零的交互效应，而且交互效应的真实估计值也在 100 次 Bootstrap 重抽样方法构造的置信区间内，从而再一次从统计推断的角度说明了 DPD-ADPLASSO 稳健变量选择估计量结果的稳定性和变量选择的一致性。

附录 2 同时展示了当离群比例增加为 0.2 时基于 Bootstrap 构建的β_j和$\gamma_{jj'}$的置信区间，可以对比发现，虽然第一个到第六个自变量的估计值的置信区间确实包含了真实的主效应系数值，也即准确选择出了主效应，但交互效应只选出了前五个，最后 X5 和 X6 变量的交互并未选出，这也和上述的 4.3 节中离群情形五的模拟结果呼应。

第五章 应用研究

5.1 血浆中 β-胡萝卜素含量的影响因素分析

5.1.1 数据和变量说明

该数据来自于 Nierenberg et al. (1989)的一项流行病学研究，主要是为了考察血浆中 β-胡萝卜素的含量与人体的 BMI 指数、年龄、吸烟状况、饮酒情况、运动情况等因素之间的关系。很多流行病学研究表明血浆中的 β-胡萝卜素能提高机体的免疫力，提高细胞间隙的连接通讯，参与人体较高级别的代谢、运转及维护。还能清除体内的自由基，而这些自由基又有可能会引起癌症，因此又有预防肿瘤的重大医学作用，除此之外，血浆中的 β-胡萝卜素在心脑血管系统、增强记忆力、视觉、美容、生殖系统、延缓衰老等方面发挥重要作用。因此，研究人体血浆中 β-胡萝卜素含量与个体各个特征之间的关系，对营养学家和临床医生更有针对性地为病人提供合理的饮食建议，甚至治疗有重要意义。

该数据集包含的变量为 Age（年龄）、Gender（性别，女=0，男=1）、BMI（体质指数，kg/m2）、Calories（每天卡路里消耗量）、Fat（每天脂肪消耗量，克）、Fiber（每天纤维消耗量，克）、Alcohol（每周饮酒量）、Smokstat（吸烟状态，从未吸烟=1，曾经吸烟=2，当前吸烟=3）、Cholesterol（每天胆固醇消耗量，毫克）、Betadiet（每天饮食中 β-胡萝卜素消耗量，微克）、Retdiet（每天饮食中维生素 A 消耗量，微克）以及 Beta-carotene（血浆中 β-胡萝卜素含量，ng/ml）。其中数据样本量为 314，分析中只考虑连续型变量，响应变量是血浆中 β-胡萝卜素含量。

5.1.2 分析结果

首先绘制响应变量直方图，QQ 图以及协变量 Alcohol 直方图。从图 4.1 中的前两个图可以看出， Beta-carotene 的分布并不是严格服从正态分布，存在离群点。并且解释变量年龄也存在部分离群点，因此将 DPD-ADPLASSO 稳健变

量选择方法应用到该数据集中。

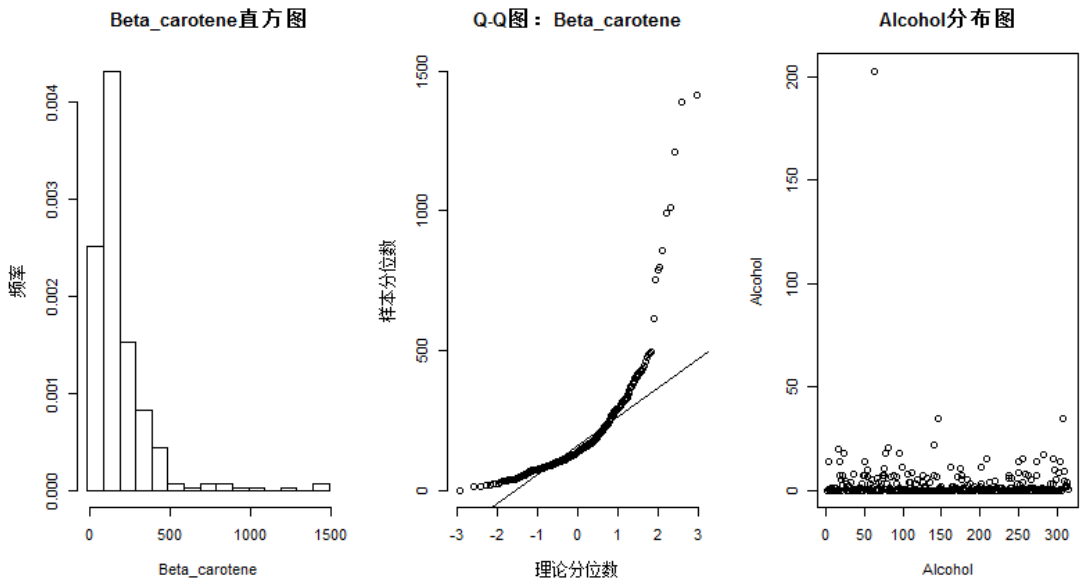

图 5.1 Beta-carotene 直方图，QQ 图以及 Alcohol 直方图

其次，对数据的解释变量找出相关系数矩阵，详见图 5.2。图中的横轴和纵轴意义相同，也即按顺序依次为代表了 9 个主效应和 36 个交互效应，所以构成了 45×45 的相关系数矩阵。通过图 5.2 可以看出，不同的解释变量之间呈现带状（banded）的结构，与 4.1 节的离群情形四的设置比较相似。

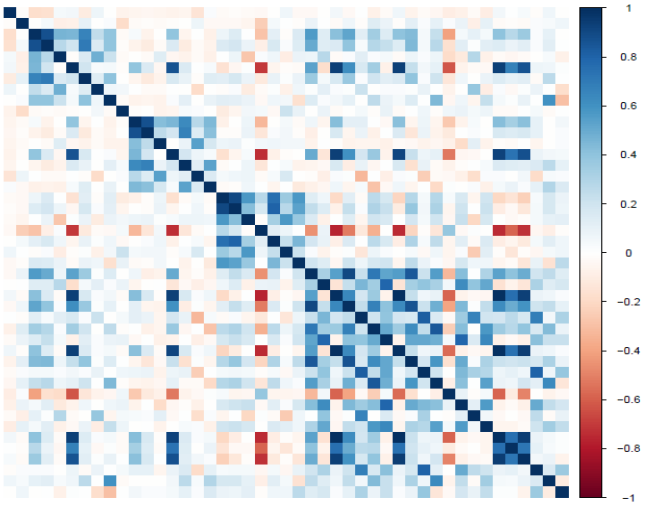

图 5.2 所有解释变量图相关系数矩阵

接着，为了验证实际数据满足强分层结构，计算每个主效应和交互效应的单

变量线性回归系数，并将其展现在图 5.3 中，图 5.3 与 5.2 的横纵轴表示意义相同，即横轴和纵轴数字 1 到 9 依次代表 9 个主效应，也即图 5.3 中的对角线，值的大小参照图中右侧的标尺，颜色越深，绝对值越大。通过下图可以看出，关于蓝色的部分，Calories 和 Alcohol 对响应变量的影响较大，并且 Calories 和 Alcohol 的交互作用对响应变量的影响也相对较大，同理，观察红色部分，BMI 和 Beta-diet 各自及其交互效应对响应变量的影响也相对较大。因此可以判断该数据满足强分层结构。因此对于满足强分层结构并且响应变量存在离群值的情况，为了更稳健的参数估计和变量选择，将本文提出的基于强分层的 DPD-ADPLASSO 应用到此数据中去。

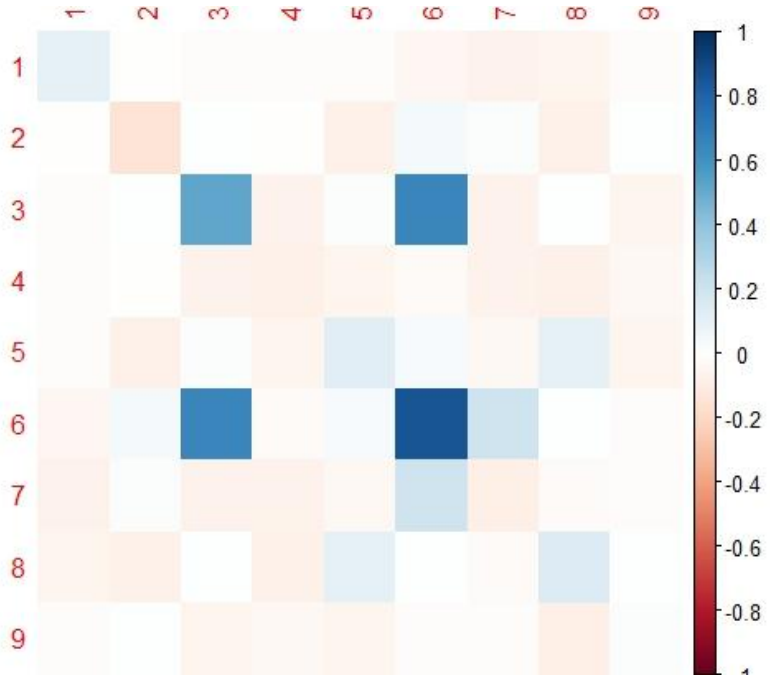

图 5.3 所有解释变量单变量回归系数图

Liu, Wang & Liang (2011)就已对该数据以 Beta-carotene 为响应变量建立可加部分线性模型进行分析并对参数部分进行变量选择，本研究在此基础之上，进一步考虑解释变量之间的交互效应，并对参数进行估计，同时实现变量选择，且保证模型的强分层结构成立。响应变量 Beta-carotene，参数部分解释变量包括 Age、BMI、Calories、Fat、Fiber、Alcohol、Betadiet、Cholesterol 和这 9 个连续型变量两两之间的交互效应，因此共有 45 个解释变量。关于最优调整参数

t，本文选取，本文参照对应最大有效性的结果作为最终的估计结果(Wang et al., 2013)的方法。为了对比，除了用本文的方法，分别还用 LAD-ADPLASSO, MM 方法估计解释变量的参数估计值，主效应的估计值如下表 5.1，交互效应的估计值请见附录。

分析结果显示，从表 5.1 中可以看出 DPD-ADPLASSO 比 LAD-ADPLASSO 和 MM 方法的变量选择出的非零主效应少，但选出来的三个解释变量的系数符号与另外两种方法的估计值符号一致。可以看出 LAD-ADPLASSO 和 MM 方法主效应的系数估计值均非零，此外从附录可以看到，LAD-ADPLASSO 选出了交互效应系数为零的变量，MM 方法估计出的交互效应系数值均为非零，而 DPD-ADPLASSO 方法的结果全为零。从而说明在本实例中 DPD-ADPLASSO 相比于 LAD-ADPLASSO 和 MM 方法，估计得到了更加稀疏的模型。

并且共同选出的变量有 BMI, Fiber 和 Beta-diet，也即影响人体血浆中 β-胡萝卜素含量的因素主要有体质指数，每天纤维素消耗量以及饮食中 β-胡萝卜素消耗量。每天纤维素消耗量，以及饮食中 β-胡萝卜素消耗量对血浆中 β-胡萝卜素含量具有正向作用，而体质指数 BMI 的值对血浆中 β-胡萝卜素含量具有负向作用。

为了查看 DPD-ADPLASSO 稳健变量选择的准确性和预测效果，随机选取原始数据集中的 200 个样本作为训练集合，其余样本作为测试集合，重复抽样 100 次，再来比较不同种方法的预测效果。从图 5.3 的预测结果箱线图可以看出，DPD-ADPLASSO 预测误差稍优于 LAD-ADPLASSO，其上四分位接近于 LAD-ADPLASSO 的下四分位数，二者预测效果总体上比较接近。而 DPD-ADPLASSO 预测效果要比 MM 方法要好得多。无论从模型的稀疏性还是从预测效果来看， DPD-ADPLASSO 在该数据集中表现效果都不错。

表 5.1　β-胡萝卜素数据：主效应协变量系数估计值

协变量	DPD 系数估计值	LAD 系数估计值	MM 系数估计值
Age	0	15.044	14.513
BMI	-23.257	-23.211	-29.331
Calories	0	-6.941	-30.647
Fat	0	-10.464	2.258
Fiber	17.17	13.917	18.681
Alcohol	0	-14.761	-38.236
Cholesterol	0	5.043	6.176
Beta-diet	12.871	15.131	20.653
Ret-diet	0	14.436	10.686

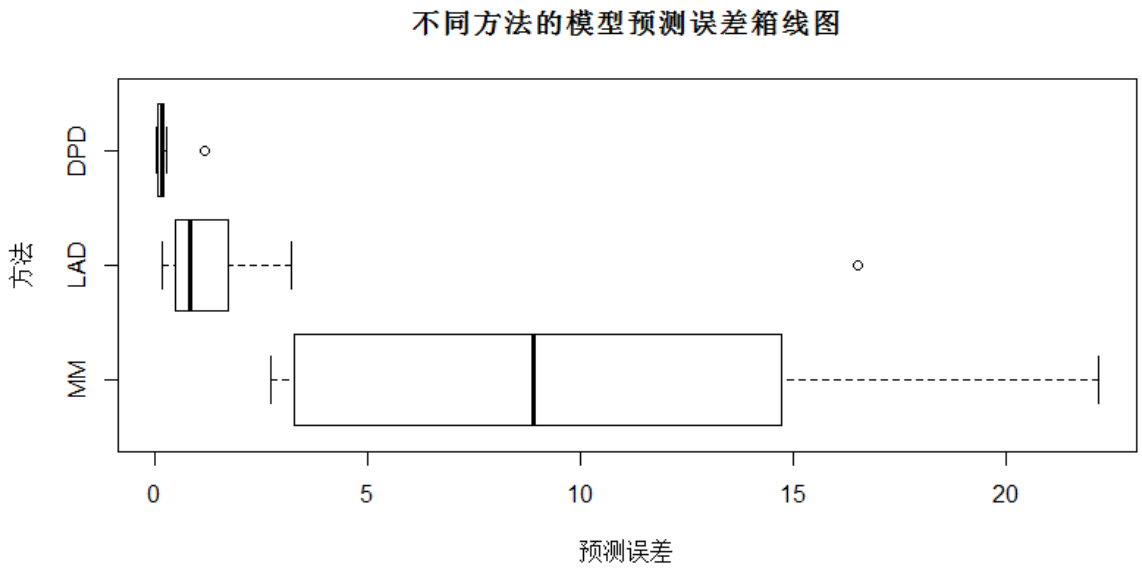

图 5.4　β－胡萝卜素含量的影响因素不同方法预测结果箱线图

注：DPD 表示 DPD-ADPLASSO 方法，LAD 表示 LAD-ADPLASSO 方法

5.2 红酒质量的识别因素分析

5.2.1 数据和变量说明

该分析数据来源于 UCI 机器学习数据库（UCI Repository of Machine Learning Databases），是关于印度地区红酒质量的化学分析，研究目的主要是探索红酒质量的影响因素。响应变量为 Quality（红酒质量），解释变量包括 Fixed-acidity（非游离酸含量）、Volatile acidity（挥发性酸含量）、Citric acidity（柠檬酸含量）、Residual sugar（残留糖含量）、Chlorides（氯化物含量）、Free sulfur dioxide（游离二氧化硫含量）、Total sulfur dioxide（总计二氧化硫含量）、Density(密度)、PH 值（酸碱度）、Sulphates（硫酸盐含量）、Alcohol(酒精)，均为连续型。总的来说参数部分解释变量包括 Fixed-acidity、Volatile acidity、Citric acidity、Residual sugar、Chlorides、Free sulfur dioxide、Total sulfur dioxide、Density、Sulphates、PH 值、Alcohol 和这 11 个连续型变量两两之间的交互效应，因此共有 66 个解释变量。研究数据集无缺失值，最终纳入分析的样本量为 1599。

首先绘制响应变量及解释变量的直方图，发现红酒质量的分布并不是严格服从正态分布，存在离群点。并且解释变量硫酸盐含量也存在部分离群点，因此将 DPD-ADPLASSO 稳健变量选择方法应用到该数据集中。

其次，要判断数据是否满足强分层的条件，因为本文的主要研究目的是探究满足强分层的前提假设下再采用 DPD-ADPLASSO 稳健变量选择的方法。同 5.1 节，再次绘制出各个解释变量的回归系数图，见图 5.5。发现挥发性酸含量和总计二氧化硫含量的主效应系数颜色较深，这两个变量的交互效应系数不为零，但颜色较弱。因此不能足够判定该数据满足强分层的假设条件。但是依据 4.3 节的模拟结果表 4.6，结合本数据中解释变量是相互独立的结构，因此强分层和不分层的结果可以近似保持一致，并且强分层的预测精度也要优于不分层的预测准确度。也即最终可以将本文的研究方法，即基于强分层的 DPD-ADPLASSO 稳健变量选择方法应用到该数据中。

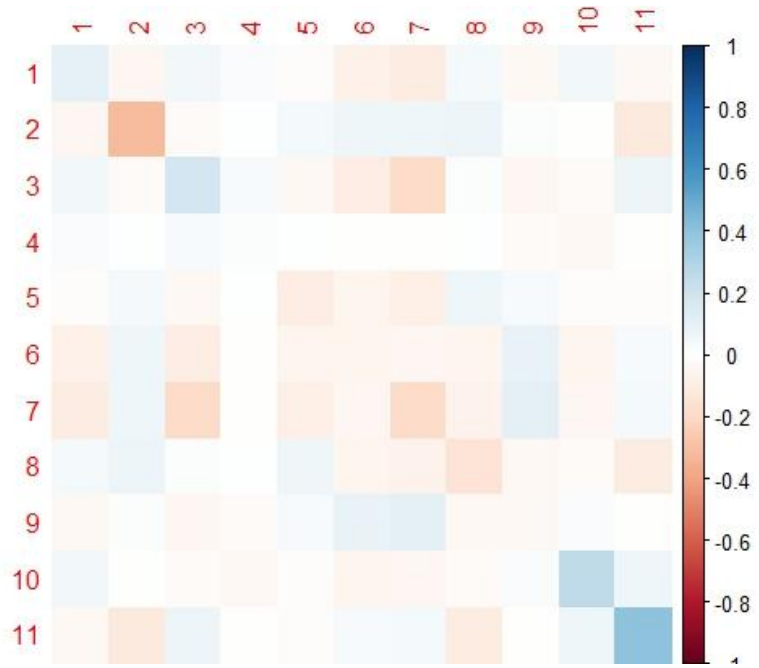

图 5.5 所有解释变量单变量回归系数图

5.2.2 分析结果

关于最优调整参数 t，本文仍然参照对应最大有效性的结果作为最终的估计结果的方法。为了对比，除了用本文的方法，再次分别用 LAD-ADPLASSO, MM 方法估计解释变量的参数估计值，主效应的估计值如下表 5.2，交互效应的估计值请见附录。

从表 5.2 可以看出，MM 估计方法并没有实现变量选择，只是做出了参数的估计，在 DPD-ADPLASSO 与 LAD-ADPLASSO 估计结果上，可以看出，两者估计出非零的主效应变量是挥发性酸含量，氯化物含量，总计二氧化硫含量，硫酸盐和酒精含量，并且这些变量的符号都保持一致，其中从解释意义上来说，挥发性酸含量，氯化物含量，总计二氧化硫含量越高红酒质量越差；硫酸盐和酒精含量越高，红酒质量越好。从附录的交互效应估计结果也可以看出挥发性酸含量与总计的二氧化硫含量对红酒质量起到了正向的作用。

为了查看 DPD-ADPLASSO 稳健变量选择的准确性和预测效果，随机选取原始数据集中的 1000 个样本作为训练集合，其余样本作为测试集合，重复抽样 100 次，再来比较不同种方法的预测效果。从图 5.6 的预测结果箱线图可以看出，

从中位数的比较来看，DPD-ADPLASSO 预测误差要优于 LAD-ADPLASSO，而且 DPD-ADPLASSO 预测效果要比 MM 方法要好得多。因此无论从模型的稀疏性还是从预测效果来看， DPD-ADPLASSO 在该非医学的数据集中表现效果也不错。

表 5.2　红酒质量数据：主效应协变量系数估计值

协变量	DPD 系数估计值	LAD 系数估计值	MM 系数估计值
Fixed acidity	0	0.026	0.05
Volatile acidity	-0.172	-0.126	-0.157
Citric acid	0	0	-0.069
Residual sugar	0	0	0.024
chlorides	-0.038	0	-0.076
Free sulfur dioxide	0	0	0.093
Total sulfur dioxide	-0.071	-0.057	-0.184
Density	0	0	-0.024
pH	-0.023	0	-0.048
Sulphates	0.176	0.138	0.224
Alcohol	0.308	0.381	0.279

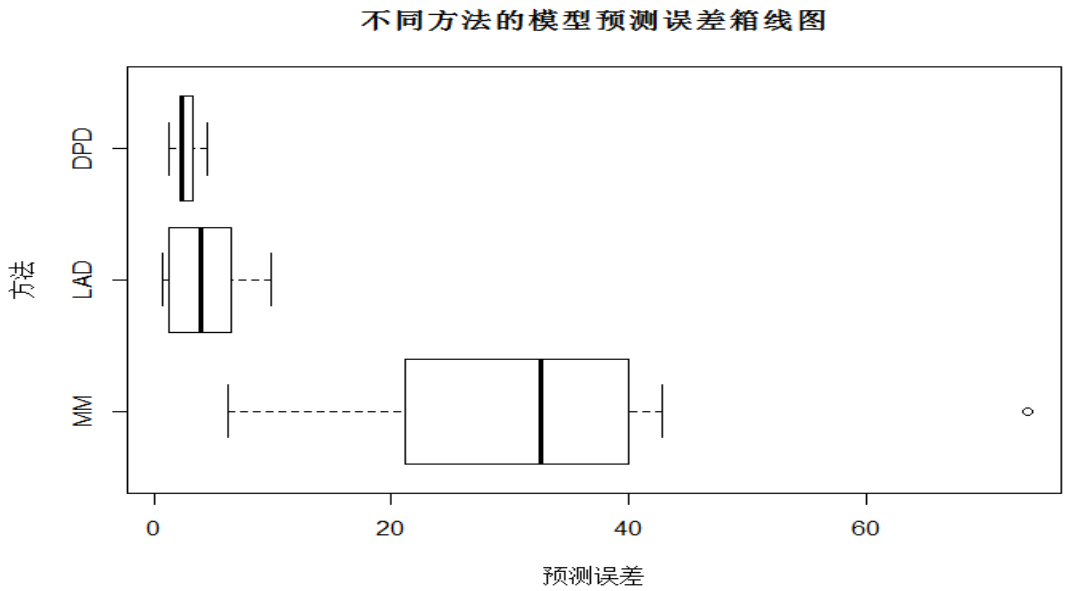

图 5.6 红酒质量的影响因素不同方法预测结果箱线图

注：DPD 表示 DPD-ADPLASSO 方法，LAD 表示 LAD-ADPLASSO 方法

第六章 总结与展望

6.1 总结

本文在系统总结现有研究文献的基础上，围绕带有交互效应的普通线性模型的稳健变量选择为主题，结合 DPD 的损失函数和自适应 LASSO 的罚函数来研究模型的稳健变量选择及相关问题，并且保证交互效应的强分层结构成立。本文在证明 DPD-ADPLASSO 稳健的变量选择在模型估计量的一致性性质的基础上，详细探讨了求解参数惩罚估计的块坐标梯度下降算法，且利用 BIC、最小化模型误差等准则来确定最优的调整参数。

为验证本文对带有交互效应的线性模型稳健估计的变量选择方法效果，本文在不同样本量、不同离群点情形及比例和不同的解释变量相关结构下进行了详细的模拟研究，且与 LAD-ADPLASSO 稳健变量选择方法相对比，比较了三种不同的方法在变量选择、参数估计和模型预测各方面的表现。模拟结果显示在模型误差预测方面，DPD-ADPLASSO 方法的效果远远优于 LAD-ADPLASSO 方法；在灵敏度以及特异度方面，在样本量较小时灵敏度与 LAD-ADPLASSO 相差不大，但随着样本量增大，灵敏度，特异度和总体分类准确度，相对于 LAD-ADPLASSO 方法效果要好。此外，当解释变量的相关结构越来越强时，模型的估计和预测准确度强分层要优于不分层的结果。这可以为实际工作者在具体问题中进一步筛选协变量提供参考。

另外，本文分别对 β-胡萝卜素数据和红酒质量数据建立交互效应模型进行稳健变量选择研究，分析了血浆中 β-胡萝卜素含量及红酒质量的影响因素，表明了本文研究方法在实际问题中，具有可观的应用效果和现实可行性。

6.2 研究的不足与展望

由于时间与知识等方面的限制，本文的研究尚存在许多不足之处，需要在接下来的研究中进一步探讨，主要问题包括如下四点：

(1) 在实际应用中，在研究个体患某种疾病的危险因素时，通常可以测量得到大量基因与环境等有关变量的信息，也即协变量具有高维性（high dimensionality），但是本文只限于低维情形，但是如果推广到高维情形或者广义线性模型，会有更强的实际使用价值。

(2) 本文研究中涉及到多个参数的选择，如惩罚函数中对参数$\boldsymbol{\beta}$、$\boldsymbol{\gamma}$的调整参数λ_β、λ_γ，这些参数的合理选择对模型拟合与变量选择结果都有重要影响。本文为简便计，令$\lambda_\beta = \lambda_\gamma$，取为一个定值，虽然模拟结果整体上较好，但也存在一些问题，例如当样本量很小时，特异度的表现略低于 LAD-ADPLSSO，这在一定程度上可以通过合理选取不同的调整参数λ_β、λ_γ来解决。基于具体数据对这些参数的合理选择将使得本文的研究方法具有更大的适用性。

(3) 本文的交互效应只简单地认为是相应主效应相乘的形式，这一般只适合于连续型变量，而对于连续型变量和分类型变量的交互效应、分类型变量和分类型变量的交互效应的具体形式及其分层变量选择问题，还需要在本文研究的基础上做出进一步探讨。

(4) 本文新提出的算法并未从理论角度上证明出 DPD-ADPLSSO 稳健变量选择具有 Oracle 性质，由于在证明中需要考虑到新提出的算法对理论性质的影响，文中只证明出了估计量具有一致性，Oracle 性质的证明需要未来进一步的研究。

附录

1．估计量理论性质证明之参数估计量的一致性

不失一般性，假定模型中β_j^*和$\alpha_{jj'}^*$代表真实非零的参数，并假定模型满足强分层结构：当$\beta_j^* = 0$或$\beta_{j'}^* = 0$时，有$\alpha_{jj'}^* = 0$。令$\boldsymbol{\theta}^* = （\boldsymbol{\beta}^{*T}, \boldsymbol{\gamma}^{*T}）^T$，$\boldsymbol{\theta}$为所有的未知参数。其中

$$\gamma_{jj'}^* = \begin{cases} \dfrac{\alpha_{jj'}^*}{\beta_j^* \beta_{j'}^*} & \beta_j^* \neq 0，\ \beta_{j'}^* \neq 0， \\ 0 & 其他 \end{cases}$$

令：

$$\mathcal{A}_1 = \{j: \beta_j^* \neq 0\}$$

$$\mathcal{A}_2 = \{(k,k'): \gamma_{kk'}^* \neq 0\}$$

$$\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2$$

$$a_{\boldsymbol{n}} = \max\{\lambda_j^{\beta}, \lambda_{kk'}^{\gamma}: j \in \mathcal{A}_1, (k,k^{'}) \in \mathcal{A}_2\}$$

$$b_{\boldsymbol{n}} = \max\{\lambda_j^{\beta}, \lambda_{kk'}^{\gamma}: j \in \mathcal{A}_1^c, (k,k^{'}) \in \mathcal{A}_2^c, k,k^{'} \in \mathcal{A}_1\}$$

$$\mathrm{H}(\theta，\ \mathrm{T}，\ \rho) = \mathrm{T}\frac{2}{\rho}\int \boldsymbol{x}\boldsymbol{x}^T e^{-\frac{r^2}{\rho}}(\frac{2r^2}{\rho} - 1) dF(\boldsymbol{X}, Y)$$

其中$\mathrm{F(X,Y)}$为解释变量和响应变量的联合分布函数。再定义如下的正则性条件：

（C1）随机误差项在零点连续且具有正的密度函数；

（C2）矩阵$\mathrm{E(XX^T)}$存在且正定；

（C3）$\mathrm{E}\|\mathrm{X}\|^3 < \infty$

条件（C1）和（C2）是为了保证 LAD 估计结果具有一致性和渐近正态性。条件（C3）是为了目标函数泰勒展开的前两项具有支配作用，可以控制展开式的余项。

定理 1.一致性

假定条件（C1）和（C2）满足，当$n\to\infty$，存在σ_*^2，满足$H(\theta^*，T（\sigma_*^2），\rho（\sigma_*^2）)$负定，且$\sigma_n^2-\sigma_*^2=o(1)$，$b_n=o(1)$，那么DPD+ADPLASSO的类型估计量满足$\|\theta_n^*-\theta^*\|=O(n^{-1/2}+a_n)$。

定理1证明如下：

令$\eta_n=n^{-1/2}+a_n$，即需要证明$\|\theta_n^*-\theta^*\|=O(\eta_n)$。等价于证明，对于任意给定的$\varepsilon>0$，存在较大的常数C，使得

$$p\left(\inf_{\|\delta\|=C} f_\alpha(\theta^*+\eta_n\delta|\sigma_n^2)>f_\alpha(\theta^*|\sigma_n^2)\right)\geq 1-\varepsilon$$

其中δ是与θ^*维数相同的向量。上式说明DPD+ADPLASSO类型的罚函数的估计量会以$1-\varepsilon$的概率落入$\{\theta^*+\eta_n\delta：\|\delta\|=C\}$的球体内。将$f_\alpha(\theta^*+\eta_n\delta|\sigma_n^2)$在$\theta^*$处泰勒展开，可得：

为了方便，令$G_n(\theta，T，\rho)=T\sum_{i=1}^n \exp\{-(y_i-X_i\theta)^2/\rho\}\frac{2(y_i-X_i\theta)}{\rho}X_i$，

$$\begin{aligned}
&f_\alpha(\theta^*+\eta_n\delta|\sigma_n^2)-f_\alpha(\theta^*|\sigma_n^2)\\
&\leq \eta_n G_n(\theta，T，\rho)^T\delta-\frac{1}{2}\delta^T[-H(\theta^*，T（\sigma_*^2），\rho（\sigma_*^2）)]\\
&\times\delta n{\eta_n}^2\{1+o(1)\}\\
&-\sum_{j=1}^{s}\left[n\eta_n\lambda_j^\theta sign(\theta^*)\delta_j+n\eta_n^2\lambda_j^\theta\delta_j^2(1+o(1))\right]\\
&=\eta_n(G_n(\theta，T，\rho)+o(\sqrt{n}))^T\delta\\
&-\frac{1}{2}\delta^T[-H(\theta^*，T（\sigma_*^2），\rho（\sigma_*^2）)]\times\delta n{\eta_n}^2\{1+o(1)\}\\
&-\sum_{j=1}^{s}\left[n\eta_n\lambda_j^\theta sign(\theta^*)\delta_j+n\eta_n^2\lambda_j^\theta\delta_j^2(1+o(1))\right]
\end{aligned}$$

由中心极限定理可得，$n^{-1/2}(G_n(\theta，T，\rho)=O(1)$，因此上述公式的第一项等于$O(n^{1/2}\eta_n)=0（n\eta_n^2）$，选择一个充分大的C，等号右边第二项相对于第一项占主要作用，第三项有界，又因为$b_n=o(1)$，所以第二项相比于第三项也占主要作用。所以总体的符号由第二项决定。因为H是负定的，所以-H正定，因此对于充分大的常数C，等号右边恒小于0。所以$f_\alpha(\theta^*+\eta_n\delta|\sigma_n^2)<f_\alpha(\theta^*|\sigma_n^2)$几

乎处处成立。从而$\sqrt{n}$ 一致性得证。

2.

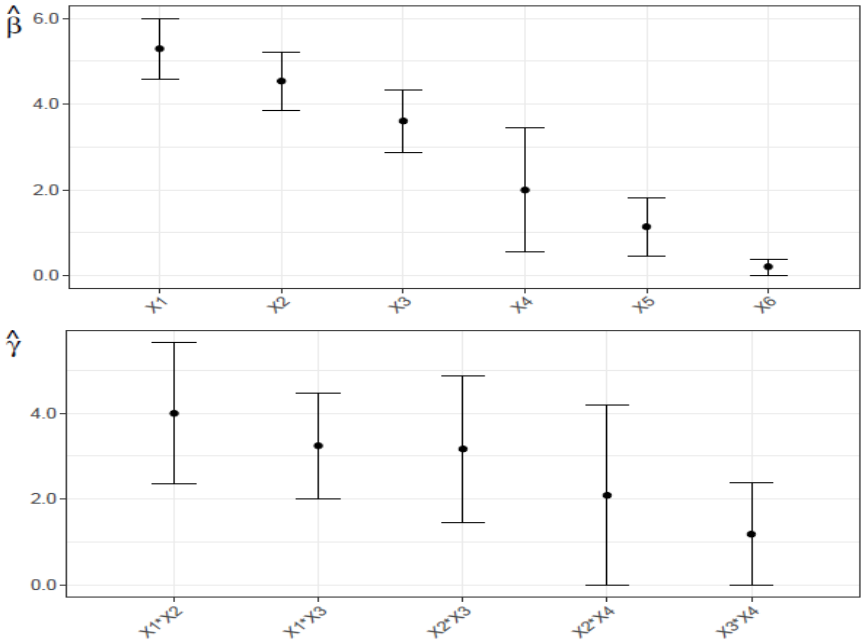

图 4.5 补充 离群比例为 0.2 时基于 Bootstrap 构建的β_j和$\gamma_{jj'}$的置信区间

3. 5.1 节血浆中 β-胡萝卜素含量的影响因素分析主效应和交互效应的估计值如下：

表 5.1s β-胡萝卜素含量的影响因素分析主效应和交互效应的估计值

协变量	DPD 系数估计值	LAD 系数估计值	MM 系数估计值
Age	0	15.044	14.513
Bmi	-23.257	-23.211	-29.331
Calories	0	-6.941	-30.647
Fat	0	-10.464	2.258
Fiber	17.17	13.917	18.681
Alcohol	0	-14.761	-38.236
Cholesterol	0	5.043	6.176
Betadiet	12.871	15.131	20.653
Retdiet	0	14.436	10.686
Age*Bmi	0	0	-7.059
Age*Calories	0	33.297	9.153

Age*Fat	0	0	5.116
Age*Fiber	0	-2.76	-9.718
Age*Alcohol	0	-31.211	-21.436
Age*Cholesterol	0	-25.333	-17.647
Age*Betadiet	0	-2.068	6.791
Age*Retdiet	0	-3.492	0.428
Bmi*Calories	0	44.958	44.352
Bmi*Fat	0	-43.524	-44.519
Bmi*Fiber	0	-18.517	-15.941
Bmi*Alcohol	0	-7.477	-0.793
Bmi*Cholesterol	0	10.119	15.442
Bmi*Betadiet	0	-5.616	-3.77
Bmi*Retdiet	0	6.545	4.01
Calories*Fat	0	55.653	30.477
Calories*Fiber	0	30.062	24.279
Calories*Alcohol	0	87.397	772.065
Calories*Cholesterol	0	-56.851	-69.503
Calories*Betadiet	0	-39.06	-6.119
Calories*Retdiet	0	-5.449	24.953
Fat*Fiber	0	-46.495	-35.374
Fat*Alcohol	0	-63.719	-224.651
Fat*Cholesterol	0	0	22.191
Fat*Betadiet	0	16.183	-17.991
Fat*Retdiet	0	-1.135	-21.855
Fiber*Alcohol	0	-18.336	-1.372
Fiber*Cholesterol	0	33.73	40.12
Fiber*Betadiet	0	27.719	10.269
Fiber*Retdiet	0	-2.084	-12.834
Alcohol*Cholesterol	0	-13.455	-34.73
Alcohol*Betadiet	0	28.472	20.622
Alcohol*Retdiet	0	7.929	-23.351
Cholesterol*Betadiet	0	0	8.318
Cholesterol*Retdiet	0	0	-16.65
Betadiet*Retdiet	0	-6.467	6.786

3. 5.2 节红酒质量的影响因素分析主效应和交互效应的估计值如下：

表 5.2s 红酒质量数据：主效应和交互效应协变量系数估计值

协变量	DPD 系数估计值	LAD 系数估计值	MM 系数估计值
fixed.acidity	0	0.026	0.05
volatile.acidity	-0.172	-0.126	-0.157
citric.acid	0	0	-0.069
residual.sugar	0	0	0.024
chlorides	-0.038	0	-0.076
free.sulfur.dioxide	0	0	0.093
total.sulfur.dioxide	-0.071	-0.057	-0.184
density	0	0	-0.024
pH	-0.023	0	-0.048
sulphates	0.176	0.138	0.224
alcohol	0.308	0.381	0.279
fixed.acidity*volatile.acidity	0	0	-0.125
fixed.acidity*citric.acid	0	0	-0.105
fixed.acidity*residual.sugar	0	0	-0.008
fixed.acidity*chlorides	0	0	-0.118
fixed.acidity*free.sulfur.dioxide	0	0	-0.059
fixed.acidity*total.sulfur.dioxide	0	0	0.052
fixed.acidity*density	0	0	0.02
fixed.acidity*pH	0	0	0.073
fixed.acidity*sulphates	0	0	0.102
fixed.acidity*alcohol	0	0	-0.085
volatile.acidity*citric.acid	0	0	0.007
volatile.acidity*residual.sugar	0	0	-0.042
volatile.acidity*chlorides	0	0	0.027
volatile.acidity*free.sulfur.dioxide	0	0	-0.031
volatile.acidity*total.sulfur.dioxide	0.059	0.035	0.116
volatile.acidity*density	0	0	0.12
volatile.acidity*pH	0	0	-0.066
volatile.acidity*sulphates	0	0	-0.012
volatile.acidity*alcohol	0	0	0.119
citric.acid*residual.sugar	0	0	-0.024
citric.acid*chlorides	0	0	0.021
citric.acid*free.sulfur.dioxide	0	0	0.037
citric.acid*total.sulfur.dioxide	0	0	0.018

citric.acid*density	0	0	0.107
citric.acid*pH	0	0	-0.134
citric.acid*sulphates	0	0	-0.034
citric.acid*alcohol	0	0	0.135
residual.sugar*chlorides	0	0	-0.018
residual.sugar*free.sulfur.dioxide	0	0	0.018
residual.sugar*total.sulfur.dioxide	0	0	-0.007
residual.sugar*density	0	0	0
residual.sugar*pH	0	0	0.009
residual.sugar*sulphates	0	0	0.019
residual.sugar*alcohol	0	0	-0.007
chlorides*free.sulfur.dioxide	0	0	-0.025
chlorides*total.sulfur.dioxide	0	0	-0.044
chlorides*density	0	0	0.082
chlorides*pH	0	0	-0.027
chlorides*sulphates	0	0	-0.018
chlorides*alcohol	0	0	0.002
free.sulfur.dioxide*total.sulfur.dioxide	0	0	-0.039
free.sulfur.dioxide*density	0	0	0.104
free.sulfur.dioxide*pH	0	0	-0.036
free.sulfur.dioxide*sulphates	0	0	-0.056
free.sulfur.dioxide*alcohol	0	0	0.091
total.sulfur.dioxide*density	0	0	-0.149
total.sulfur.dioxide*pH	0	0	0.064
total.sulfur.dioxide*sulphates	0	0	0.033
total.sulfur.dioxide*alcohol	0	0	-0.172
density*pH	0	0	0.028
density*sulphates	0	0	-0.093
density*alcohol	0	0	0.003
pH*sulphates	0	0	0.063
pH*alcohol	0	0	-0.007
sulphates*alcohol	0	0	0.019

参考文献

1. Yajima I, Kumasaka MY, Naito Y, Yoshikawa T, Takahashi H, et al. (2012) Reduced GNG2 expression levels in mouse malignant melanomas and human melanoma cell lines. Am J Cancer Res 2: 322-329.
2. Shields PG (2002) Molecular epidemiology of smoking and lung cancer. [J]Oncogene 21: 6870-6876
3. McCullagh P, Nelder J A. Generalized linear models[M]. London: Chapman and Hall, 1989.
4. McCullagh P. What is a statistical model[J]. The Annals of statistics, 2002, 30: 1225-1267.
5. Chipman H. Bayesian variable selection with related predictors[J]. Canadian Journal of Statistics, 1996, 24(1): 17-36.
6. Chipman H, Hamada M, Wu C F J. A Bayesian variable-selection approach for analyzing designed experiments with complex aliasing[J]. Technometrics, 1997, 39(4): 372-381.
7. Bien J, Taylor J, Tibshirani R. A lasso for hierarchical interactions[J]. The Annals of Statistics, 2013, 41(3): 1111-1141.
8. Akaike H. Information theory and an extension of the maximum likelihood principle. 2nd International symposium on information theory. Budapest: Akademiai Kiado. 1973, 267-281.
9. Schwarz G. Estimating the dimension of a model[J]. The Annals of Statistics, 1978, 6: 461-464.
10. Mallows C L. Some comments on Cp[J]. Technometrics, 1973, 15(4): 661-675.
11. Tibshirani R, Knight K. The covariance inflation criterion for adaptive model selection[J].Journal of Royal Statistical Society: Series B (Statistical Methodology),1999, 3:529-546.
12. Breiman L. Better subset regression using the nonnegative garrote[J]. Technometrics, 1995, 37(4): 373-384.
13. Tibshirani R. Regression shrinkage and selection via the lasso[J]. Journal of the Royal Statistical Society. Series B (Methodological), 1996, 58: 267-288.
14. Knight K, Fu W. Asymptotics for lasso-type estimators[J]. The Annals of

Statistics, 2000, 28: 1356-1378.
15. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2008, 70(5): 849-911.
16. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties[J]. Journal of the American Statistical Association, 2001, 96 1348-1360.
17. Efron B, Hastie T, Johnstone I, et al. Least angle regression[J]. The Annals of statistics, 2004, 32(2): 407-499.
18. Zou H, Hastie T. Regularization and variable selection via the elastic net[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2005, 67: 301-320.
19. Zou H. The adaptive lasso and its oracle properties[J]. Journal of the American statistical association, 2006, 101: 1418-1429.
20. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2006, 68(1): 49-67.
21. Zhang C H. Nearly unbiased variable selection under minimax concave penalty[J]. The Annals of Statistics. 2010, 38: 894-942.
22. SEBER, LEE. Linear Regression Analysis[M]. 2nd ed. New York : Wiley, 2003
23. HUBER P J. Robust regression: Asymptotics, Conjectures and Monte Carlo[J]. The Annals of Statistics, 1973, 1 : 799 – 821.
24. ROUSSEEUW P, YOHAI V. Robust Regression by means of S-Estimators[J]. Robust and Nonlinear Times Series, 1984, 26 : 256 – 272.
25. YOHAI V J, ZAMAR R H. High breakdown-point estimates of regression by means of minimization of an efficient scale[J]. Journal of the American Statistical Association, 1988, 83 : 406 – 413.
26. WANG H, LI G, JIANG G. Robust Regression Shrinkage and Consistent Variable Selection Through the LAD-Lasso[J]. Journal of Business and Economic Statistics, 2007, 25 : 347 – 355.
27. LENG C. Variable Selection and Coefficient Estimation via Regularized Rank Regression[J]. Statistica Sinica, 2010, 20 : 167 – 181.

28. Wang, X., Jiang, Y., Huang, M., Zhang, H.(2013) Robust variable selection with exponential squared loss. Journal of the American Statistical Association. 108:632–643.
29. BASU A, HARRIS I R, HJORT N L, et al. Robust and efficient estimation byminimizing a density power diver-gence[J]. Biometrik, 1998, 85(3) : 549 – 559.
30. DURIO A, ISIIA E D. The Minimum Density Power Divergence Approach in Building Robust Regression Models[J]. INFORMATICA, 2011, 22. No.1(1) : 43 – 56.
31. GHOSH A, BASU A. Robust Estimation for Independent Non-Homogeneous Observations using Density Power Divergence with Applications to Linear Regression[J]. Electronic Journal of statistics, 2013a, 7 : 2420 – 2456.
32. 李根，邹国华，张新雨.高维模型选择方法综述[J].数理统计与管理，2012,4:640-656
33. Yuan M, Joseph V, Lin Y. An efficient variable selection approach for analyzing designed experiments[J]. Technometrics, 2007, 49(4): 430-439.
34. Yuan M, Joseph V, Zou H. Structured variable selection and estimation[J]. The Annals of Applied Statistics, 2009, 3(4): 1738-1757.
35. Zhao P, Rocha G, Yu B. The composite absolute penalties family for grouped and hierarchical variable selection[J]. The Annals of Statistics, 2009, 37: 3468-3497.
36. Radchenko P, James G M. Variable selection using adaptive nonlinear interaction structures in high dimensions[J]. Journal of the American Statistical Association, 2010, 105(492): 1541-1553.
37. Bien J, Taylor J, Tibshirani R. A lasso for hierarchical interactions[J]. The Annals of Statistics, 2013, 41(3): 1111-1141.
38. Lim M, Hastie T. Learning interactions through hierarchical group-lasso regularization[J]. arXiv preprint arXiv:1308.2719, 2013.
39. Park M Y, Hastie T. Penalized logistic regression for detecting gene interactions[J]. Biostatistics, 2008, 9(1): 30-50.
40. Choi N H, Li W, Zhu J. Variable selection with the strong heredity constraint and its oracle property[J]. Journal of the American Statistical Association, 2010, 105(489): 354-364.

41. Zhu R, Zhao H, Ma S. Identifying Gene–Environment and Gene–Gene Interactions Using a Progressive Penalization Approach[J]. Genetic epidemiology, 2014, 38(4): 353-368.
42. Wasserman L, Roeder K. High dimensional variable selection[J]. The Annals of statistics, 2009, 37(5A): 2178-2201.
43. Meinshausen N, Meier L, Bühlmann P. P-values for high-dimensional regression[J]. Journal of the American Statistical Association, 2009, 104(488): 1671-1681.
44. Minnier J, Tian L, Cai T. A perturbation method for inference on regularized regression estimates[J]. Journal of the American Statistical Association, 2011, 106(496).
45. Chatterjee A, Lahiri S N. Bootstrapping lasso estimators[J]. Journal of the American Statistical Association, 2011, 106(494): 608-625.
46. Chatterjee A, Lahiri S N. Rates of convergence of the adaptive LASSO estimators to the oracle distribution and higher order refinements by the bootstrap[J]. The Annals of Statistics, 2013, 41(3): 1232-1259.
47. Bühlmann P. Statistical significance in high-dimensional linear models[J]. Bernoulli, 2013, 19(4): 1212-1242.
48. Zhang C H, Zhang S S. Confidence intervals for low dimensional parameters in high dimensional linear models[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2014, 76(1): 217-242.
49. Lockhart R, Taylor J, Tibshirani R J, et al. A significance test for the lasso[J]. The Annals of Statistics, 2014, 42(2): 413-468.
50. Javanmard A, Montanari A. Confidence intervals and hypothesis testing for high-dimensional regression[J]. arXiv preprint arXiv:1306.3171, 2013.
51. Van De Geer S, Bühlmann P, Ritov Y, et al. On asymptotically optimal confidence regions and tests for high-dimensional models[J]. The Annals of Statistics, 2014, 42(3): 1166-1202.
52. 许文甫.基于密度势差异的稳健变量选择研究,中国人民大学[D],2016.6
53. 曾宪斌.半参数交互效应模型的分层变量选择研究，中国人民大学[D],2015.6
54. TSENG P, YUN S. A Coordinate Gradient Descent Method for Nonsmooth Separable Minimization[J]. Mathematical Programming, 2009, 117 : 387 – 423.
55. 张晓勇，王仲君.二分法和牛顿迭代法求解非线性方程的比较及应用[J],教育

教学论坛,2013,25:139

56. Freedman D A. Bootstrapping regression models[J]. The Annals of Statistics, 1981, 9(6): 1218-1228.
57. Chatterjee A, Lahiri S N. Bootstrapping lasso estimators[J]. Journal of the American Statistical Association, 2011, 106(494): 608-625.
58. Nierenberg D W, Stukel T A, Baron J A, et al. Determinants of plasma levels of beta-carotene and retinol. American Journal of Epidemiology, 1989, 130: 511-521.
59. Liu X, Wang L, Liang H. Estimation and Variable Selection for Semiparametric Additive Partial Linear Models[J]. Statistica Sinica, 2011, 21: 1225-1248.

致 谢

三年的硕士时光即将进入尾声，依稀记得 2012 年的春天，来到人民大学的校园，还在憧憬着有一天能够在这所学校读书，后来等到梦想成真的时候，是在 2014 年秋季入学时，当时的我背着行李，小心翼翼的踏入人民大学西门，心情有点小小的激动与不安。很快，进入到易师门后，所有的学习和生活踏入正轨，面对和蔼可亲的导师，面对朝气蓬勃的同学，渐渐感觉到自己进入了一个新的平台，而且我终于要在梦想的学校慢慢成长了！转眼间，三年已经悄悄的溜走，回想这三年的时光，有过太多的酸甜苦辣，不论是学习还是生活上，都感受到了成长赐予我的强大力量。那么，在论文即将完成之际,我想衷心感谢这一路上帮助我的人。

首先,衷心感谢我的导师易丹辉教授。还记得我第一次在易老师的办公室，易老师给我举一位师姐的例子，赞扬她的数据处理能力很强，巾帼不让须眉，给我留下了非常深刻的印象，我当即找到了学习的榜样。后来，在做科研项目的时候，会有很多处理数据的课题，项目组会一起讨论一个问题直到深夜，每当这个时候，我就会想起老师曾经说过的话，鼓励大家一起加油，集思广益，直到一起解决掉这个难题。我相信这是一个自我提高的过程，不近有专业知识的收获，也有了自我认可的满足感和自豪感，而这些进而会转化为动力，鼓舞我们积极参与到下一个项目中去。在整个硕士学习期间, 我总共参与了五个大大小小的项目，易老师对我们严格要求却和蔼可亲，这也是整个硕士期间最浓墨重彩的一笔。当然也会有很多生活上的小事，比如会和我聊天让我恢复自信，慷慨借用电脑等等，在这里,我要向易老师表示由衷的感谢，谢谢您这三年来对我的细心栽培，让我没有后悔过来到这个学校，并且在人民大学这样的一个优秀的平台上，谢谢您引领我快速的前进。

感谢李扬老师，李老师是我在踏入科研大门最重要的启蒙导师。研一，李老师给了我一篇英文文章，让我将里面的方法在讨论班里和大家分享，这个过程中，发生了很多个第一次，比如第一次看英文的文献，第一次做英文的展示，第一次写超过两百行的代码……每一次都离不开老师的鼓励和支持。我甚至还

非常清晰的记得，在一次周四讨论班上，我因为自控力不足，没能展示出当前阶段的成果，老师就会当即非常严肃的指出我的不足。在自惭形秽的同时，也深深的感受到老师那种恨铁不成钢的焦急。李老师不仅在学术上指引、督促我，而且在生活上也给予了极大的帮助。研二伊始决定放弃本校硕博连读的时候，自己非常无助。李老师知道后，主动和我沟通，排解我的心理压力，后来情绪才慢慢好转，才意识到那些聊天都是发生在他的休息时间！现在回想起来，每次觉得自己做的不好的时候，李老师都会温柔的告诉我，冯倩，你一定可以的！除了深深的感动，甚至一度觉得李老师是我特别想要成为的那种人。

此外，尤其感谢许文甫师兄，在我的毕业论文陷入困境的时候，给予了非常大的支持。记得差两天就 2017 农历新年的时候，我寻求师兄的帮助，他都愿意用自己的时间来帮我。真的非常感谢！还要感谢马双鸽老师，秦祎辰老师给我的点拨，让我在毕业论文的选题方上少走很多弯路，感谢密歇根大学的朱冀教授，提供了算法启蒙，也即如何实现交互效应下的强分层变量选择，还要感谢曾宪斌师兄，赵建喜师兄和陈国勇师兄；也要谢谢林玲师妹在服务器的如何使用上，提供的无私帮助。

最后，特别感谢我的爸爸妈妈，你们虽然是普普通通的农民，但在我的眼中，你们就如同具有远见格局的教授一样，那些农村里和我一样同龄的女孩都嫁人生孩子了，但是爸爸让我的人生多了很多选择，不仅在整个小学到硕士求学期间，赋予了我最强大的学习动力，而且还鼓励我出国留学深造。虽然爸爸妈妈你们不能每时每刻在我的身边，但是你们总是想办法教导我，督促我，让我对未来充满了希望。此生能够有你们，是女儿我最大的幸运！现在硕士的道路已经走向终点，但未来科研的道路才刚刚起步。女儿开始喜欢上科研这条道路，因为这条道路是那么的安静，纯粹，还伴随着不断的尝试与成就感，并且意识到，世界上真的没有什么救世主，只要你想做，just do it，那么时间就会告诉你最好的答案。

冯倩

2017 年 4 月 20 日