学校代码	10602
学　　号	2007010671
分 类 号	O212.7
密　　级	公开

广西師範大學
GUANGXI NORMAL UNIVERSITY

硕士学位论文

高维线性模型的变量选择

Variable Selection For
High-dimensional Linear Models

学　　院：数学科学学院

专　　业：概率论与数理统计

研究方向：数理统计

年　　级：2007级

研 究 生：李玲玲

指导教师：张军舰 副教授

完成日期：2010年4月

高维线性模型的变量选择

专业名称: 概率论与数理统计

申 请 人: 李 玲 玲

指导老师: 张 军 舰 副教授

论文答辩委员会

主席: _____

委员: _____

_____

_____

_____

_____

论文独创性声明

本人郑重声明: 所提交的学位论文是本人在导师的指导下进行的研究工作及取得的成果. 除文中已经注明引用的内容外, 本论文不含其他个人或其他机构已经发表或撰写过的研究成果. 对本文的研究作出重要贡献的个人和集体, 均已在文中以明确方式标明. 本人承担本声明的法律责任.

研究生签名: 李冬冬 日期: 2010.5.27

论文使用授权声明

本人完全了解广西师范大学有关保留, 使用学位论文的规定. 广西师范大学, 中国科学技术信息研究所, 清华大学论文合作部, 有权保留本人所送交学位论文的复印件和电子文档, 可以采用影印, 缩印或其他复制手段保存论文. 本人电子文档的内容和纸质论文的内容相一致. 除在保密期间内的保密论文外, 允许论文被查阅和借阅, 可以公布 (包括刊登) 论文的全部或部分内容. 论文的公布 (包括刊登) 授权广西师范大学学位办办理.

研究生签名: 李冬冬 日期: 2010.5.27

导　师签名: 张保权 日期: 2010.5.27

高维线性模型的变量选择

年级: 2007 级　专业: 概率论与数理统计　研究方向: 数理统计
研究生: 李玲玲　导师: 张军舰

摘　要

高维数据在生物信息学、生物医学、计量经济学和机器学习等领域频繁出现，它使经典的统计方法遭受了较大的冲击，是目前统计学理论和应用研究中面临困难最多、挑战最严峻的研究领域之一。对高维数据的研究，通常情况需要假定真实模型位于一个低维空间(至少协变量维数要比样本容量低)，也就是常说的稀疏性 (sparsity) 假定。然后再按照一定的规则选择变量进行降维。通过选择变量，将对因变量根本没有影响或者影响较小的变量剔除，保留影响较大的变量，从而建立简洁的模型。

本论文在对前人已有研究成果进行综合分析的基础上，主要考察随机误差项分布未知时的高维线性模型的变量选择问题。所采用的方法和思想主要是 Fan 和 Lv(J.Fan,J.Lv. Sure independence screening for ultra-high dimensional feature space[J]. Journal of the Royal Statistical Society Series B.2008,70:849-911.) 的安全独立筛选 (Sure independent screening, 简记为 SIS) 方法和迭代安全独立筛选 (Iterative sure independent screening, 简记为 ISIS) 方法，与 Chen,Variyath 和 Abraham(J.Chen,A.M.Variyath,B. Abraham. Adjusted empirical likelihood and its properties[J].Journal of Computational and Craphical Statistics.2008,17,426-443.) 的调整经验似然 (AEL) 方法的有机结合。从理论上进一步说明 Fan 和 Lv 的 SIS 方法和 ISIS 方法在误差项不服从正态分布时也是成立的，并且具有较好的安全筛选渐近性质。即在一定条件下可以得到：

$$P(M_* \subset M_\gamma) = 1 - O(\exp(-Cn^{1-2\kappa}/\log(n))),$$

这里 M_* 为我们感兴趣的真实稀疏模型的指标集，M_γ 为自变量 X 与因变量 Y 相关性最强的前 $[n\gamma]$ 个变量所对应下标构成的指标集，其中 $\gamma \in (0,1), 1-2\kappa > 0$。由此可知，对充分大的样本容量 n，选择的变量包含真实变量的概率接近于 1。从算法上分别给出了 SIS 方法、迭代 SIS 方法与调整经验似然有机结合的 SIS+AEL 算法与迭代 SIS+AEL 算法。SIS+AEL 算法思路为：先选出与因变量相关性较强的 $[n\gamma]$ 个变量，然后由选择出来的变量与因变量利用线性模型调整经验似然比函数，得到相应的经验似然比的 AIC 和 BIC 版本，最后通过经验似然的 AIC 或 BIC 标准寻找出最优的模型，从而完成变量选择。迭代 SIS+AEL 算法思路如下：先用 SIS+AEL 方法从 $x_1, \cdots, x_p$ 中得到初步的变量 $x_{i1}, \cdots, x_{im_1}$，然后建立 y 与 $x_{i1}, \cdots, x_{im_1}$ 之间的线性模型，接着对其残差 $y - (x_{i1}\hat{\beta}_{i1} + \cdots + x_{im_1}\hat{\beta}_{im_1})$ 和剩余的 $p - m_1$ 个变量再应用 SIS+AEL 方法选择一部分变量，如此循环，直至得到我们满意的变量为止。这个算法既保留了原有方法的渐近性质，又降低了实际中对误差项的分布要求，取长补短、计算简单、想法直观。最后，我们对所提出的方法作了模拟说明。模拟结果表明，我们的算法在对高维线性模型作变量选择时，

若误差项服从的是正态分布，其包含真实模型的比率与 LASSO(现阶段比较流行的方法)相差不大；而对于误差项不服从正态分布时，其结果整体上要略好于 LASSO。

本文特色主要体现在以下几点：

1. 对已有的一些方法进行重新组合，取长补短，降低了计算的工作量，拓宽了应用的范围。

2. 去掉误差项是正态分布的限制，Fan 和 Lv(2008) 的 SIS 和 ISIS 方法尽管简单，但对误差项要求是正态分布，只有这样才能满足其相应的性质。本文从理论上说明误差项不必是正态分布，在较宽的条件下也可以得到 SIS 和 ISIS 方法相同的结论。对指标维数 p 降到样本容量 n 以下的情况，我们选择经验似然方法，无须对误差项作任何分布假定。

3. 采用调整经验似然方法作变量选择克服了经验似然的一些缺陷，众所周知，经验似然在使用时有一前提约束，即参数 θ 构造的估计方程 $E_F g(y,\theta) = 0$ 中，$\{g(y_i,\theta), i = 1,\cdots,n\}$ 的凸包必包含 0 点并且 0 为其一个内点。为避免此条件不满足时所引起的系统偏差，我们选用了 Chen,Variyath 和 Abraham(2008) 的调整经验似然方法。

关键词: 高维线性模型; 变量选择; SIS 方法; 迭代 SIS 方法; 调整经验似然

Variable Selection For High-dimensional Linear Models

Grade: 2007 Major: Probability and Statistics Research field: Statistics
Graduate: Lilingling Supervisor: Zhangjunjian

ABSTRACT

High-dimensional data frequently appear in many areas such as Bioinformatics, Biomedicine, Econometrics and Machine learning. It makes the classical statistical methods fail in most cases. So it's one of the research fields with a lot of difficulties and challenges in the statistical theories and applications. In theoretical and applied studies of the high-dimensional data, the sparsity condition is used frequently, that is, the true model lies in a low-dimensional space in which the number of the covariates is less than the sample sizes. If the sparsity condition does not be satisfied, the true model can not be indentified, then the variable selection will be meaningless. According to some criterions, we can do the varible selecion and give an approximation to the true model. In practice, we often delete the variables which are weakly correlated or uncorrelated with the response variable, and choose the variables which are highly correlated with the response.

In this thesis, the variable selection in high-dimensional linear model is investigated. We mainly consider the case that the error distribution is unknown and the dimension p is large than the sample size n. Our method is a reasonable combination between the SIS (Sure independent screening) or the ISIS(Iterative sure independent screening) in Fan and Lv(2008) and the AEL(Adjusted empirical likelihood) in Chen,Variyath and Abraham(2008). The title of Fan and Lv(2008) is " Sure independence screening for ultra-high dimensional feature space", which published in "Journal of the Royal Statistical Society Series B(70:849-911)". The other paper is "Adjusted empirical likelihood and its properties" published in the " Journal of Computational and Craphical Statistics (17,426-443)". In theoretically, we prove that the asymptotic properties of SIS and ISIS in Fan and Lv(2008) are still hold without the gaussian assumption about the error distribution. Concretely, under some conditions, we have

$$P(M_* \subset \ M_\gamma) = 1 - O(\exp(-Cn^{1-2\kappa}/\log(n))),$$

where M_* is the true sparse model; M_γ is the selected model that includs $[n\gamma]$ variables, the $[n\gamma]$ variables are highly correlated with the response variable; $\gamma \in (0,1), 1-2\kappa > 0$. The property indicates that the selected model can contain the true model in probability. In algorithmically, we give the SIS+AEL and Iterative SIS+AEL algorithms. The idea of SIS+AEL algorithm is as

follows: Firstly, we choose $[n\gamma]$ variables which are highly correlated with the response variable; Next we use the AEL to obtain the corresponding AIC and BIC, finally, the approximated model are chosen by the minimum of AIC or BIC. The iterative SIS+AEL algorithm is the following. In the first step,we use SIS+AEL to choose $x_{i1}, \cdots, x_{im_1}$ from $x_1, \cdots, x_p$, then establish a linear model with y and $x_{i1}, \cdots, x_{im_1}$. In the second step, we use SIS+AEL again to select variables, in this case, insteading of Y and $x_1, \cdots, x_p$, the response variable is the residual $y - (x_{i1}\hat{\beta}_{i1} + \cdots + x_{im_1}\hat{\beta}_{im_1})$ and the covariates are the remaining $p - m_1$ variables except $x_{i1}, \cdots, x_{im_1}$. We keep on doing this until some criterions are satisfied. The algorithms not only remain the asymptotic properties of SIS and ISIS, but also weken the assumption about the error distribution. The idea is straight and they can make up for each other's dificiencies. At last, we do some simulations. Simulation results show that if the error distribution is gaussian, the accuracy of ISIS+AEL in including the true model is close to that of LASSO(The popular method in variable selection); If the error distribution is not gaussian, the accuracy of ISIS+AEL in including the true model is better than that of LASSO.

The deserved results in the thesis are as follows:

1. The SIS , the Iterative SIS and the AEL are combined systematically. The new method can make up for each other's dificiencies and reduce the computation cost and widen the application fields.

2. The restriction about the Gaussian error distribution is removed. Although the computation of SIS and ISIS from Fan and Lv(2008) is simple, the asymptotic properties will not be hold without the gaussian error distribution. In theoretically, we prove that the asymptotic properties of SIS and ISIS in Fan and Lv(2008) are still hold under some weaker conditions. The index dimension p can reduce to $m(m < n)$ by choosing the AEL method.

3. Using the AEL method to select variables, we can overcome the shortcoming of empirical likelihood. It is well know that, there is a precondition for empirical likelihood. In the estimate equation $E_F g(y, \theta) = 0$, the estimation of θ exists if and only if the convex hull of the $\{g(y_i, \theta), i = 1, \cdots, n\}$ contains zero as an inner point. In order to avoid the systematic bias caused by the false precondition, we choice the Adjusted empirical likelihood proposed by Chen,Variyath and Abraham(2008).

Key words: High-dimensional linear model; Variable selection;Sure independent screening method;Iterative sure independent screening method;Adjusted empirical likelihood

目　录

第一章　绪论

现代技术的不断发展，许多领域都产生了海量复杂的数据，其中比较典型的是高维数据，即观测数据的指标维数比较大，它们中大部分的指标维数甚至比观测数据的个数还要大，这使得传统的统计方法遭受了较大的冲击，许多传统方法因不太适合新的数据要求而不能应用。有关方面的例子可以从生物信息学、生物医学、计量经济学和机器学习等领域很容易得到，详见 Hastie,Tibshirani 和 Friedman(2001)[1]。这些新现象产生了许多挑战性的工作，需要我们寻找更好的方法来解决。近几年国内外的大型会议几乎都有高维数据专场，并且为此专题还开展过几次国际会议。例如 2007 年于云南大学召开的高维数据分析国际会议等。因此高维数据是目前统计学理论和应用研究中面临困难最多、挑战最严峻，也是最有可能取得突破的研究领域之一。

事实上，许多高维统计学习问题都可以抽象为如下问题：从实际中可以得到一个或多个输出变量 y，以及与它们有关的特征或协变量 $x_1, \cdots, x_p$ 的 n 次观测，我们需要基于这些观测建立 y 与 $x_1, \cdots, x_p$ 的数学模型。与传统统计方法不同的是，此处一般情况协变量的维数 p 大于 n，有时甚至是远大于 $n(p \gg n)$。详见 Turlach(2004)[2]、Peter(2007)[3]等、Candes 和 Tao(2007)[4]。为建立 y 与 $x_1, \cdots, x_p$ 之间的数学模型，通常情况需要假定真实模型位于一个低维空间 (至少协变量维数 p 要比样本容量 n 低)，也就是常说的稀疏性 (sparsity) 假定，否则，所建立的模型根据所观测的样本是不可识别的。因此在维数较高时我们采取的方法一般是变量降维，也就是变量选择。变量选择是一种特殊的模型选择问题，在统计研究中有着重要的意义。在实际问题中，人们罗列出来的可能与因变量有关的自变量往往很多，其中有些变量对因变量根本没有影响或者影响很小。如果在建模过程中把这些变量都包括进来，不但计算量大，估计和预测的精度也会下降，有时甚至是不可识别的。而且，对有些实际问题，某些自变量的观测数据获得代价昂贵，如果这些自变量本来就对因变量影响不大，或者根本没有影响，而我们不加选择的将其包含进模型中，势必造成观测数据收集和模型应用的费用不必要的加大。

那么，如何对变量进行选择呢？一般情况，我们需要先对模型的类型大致有一个了解，然后再从中选择合适的模型。众所周知，线性模型是刻画 y 与协变量 $x_1, \cdots, x_p$ 之间关系的一个最基本模型。一方面，线性模型比较简单，容易计算；另一方面，许多复杂的模型可以通过变换转换为线性模型。如果能对线性模型给出较好的选择方法，很容易把此方法推广到其他模型中，因此本文主要讨论高维线性模型的变量选择问题。

本章后面部分计划如下：第一节先给出有关线性模型高维数据变量选择已有方法及研究思路，指出这些方法的优势和劣势；第二节给出我们的主要研究内容及创新之处。

§1.1 文献综述

考虑一个响应变量 y 和 p 个协变量 (或解释变量)$x_1,\cdots,x_p$ 的情况，所考虑的模型为线性模型

$$y = x\beta + \varepsilon.$$

其中 $x = (x_1,\cdots,x_p)$，$\beta = (\beta_1,\cdots,\beta_p)^T$。变量选择 (对此模型也就是模型选择) 的基本任务就是按照一定的准则从 $x_1,\cdots,x_p$ 中选出真正与 y 有关的变量 $x_{i_1},\cdots,x_{i_s}$,其中$\{i_1,\cdots,i_s\}$为$\{1,\cdots,p\}$ 的一个子集，然后建立 y 与 $x_{i_1},\cdots,x_{i_s}$ 之间的线性模型。

假设我们进行了 n 次观测，此时所考虑的模型为：

$$y_i = X_i^T\beta + \varepsilon_i, \qquad i = 1,\cdots,n, \tag{1-1}$$

其中 $X_i^T = (x_{i1},\cdots,x_{ip})$ 为协变量的第 i 次观测，$\beta = (\beta_1,\cdots,\beta_p)^T$ 是未知的回归系数向量，ε_i 为误差度量。令 $Y = (y_1,\cdots,y_n)^T$ 表示响应变量 (因变量) 的观测向量，$X = (X_1,\cdots,X_n)^T$ 是 $n \times p$ 的设计矩阵，$\varepsilon = (\varepsilon_1,\cdots,\varepsilon_n)^T$ 为误差向量。此时模型 (1-1) 可写为

$$Y = X\beta + \varepsilon. \tag{1-2}$$

关于线性模型 (1-1) 或 (1-2) 的变量选择方法比较多，特别是指标维数 p 小于 n 且 p 固定的情况，其理论相对比较成熟。但对于 p 大于 n 或 p 随 n 变化而变化的情况，尽管讨论的文献较多，但较为成熟的方法和理论相对还比较少。概括起来，目前已有的变量选择方法大致可分为四类，分别是序贯方法、预测误差方法、信息论方法以及惩罚方法。当然它们之中可能有部分是相互重合的。下面我们主要从这四个方面分别加以阐述，其他方面可参见王大荣 (2009)[5] 以及其后的参考文献。

一、序贯方法

序贯方法指的是在作变量选择时一次选择仅考察一个自变量，通过某一准则决定是否增加或删减某一自变量。经典的向前选择、向后选择和逐步回归等均属于此类。有关这方面介绍的文献较多，任一本有关多元回归分析的教材几乎都有。例如王松桂等 (2000)[6],Miller(2002)[7] 等。此处不再一一细述。这类方法的优点是理论和计算较为成熟，目前大多数统计软件均有此功能，缺点是它适合于 $p \leq n$ 且 p 固定的情况，而且这些方法的判定准则基本都是基于误差项 ε 服从正态分布假定而建立起来的，一旦误差项不服从正态分布结果可能会有所影响。

二、预测误差方法

预测误差方法指的是事先把样本分成两部分，一部分用来作变量选择，建立模型。然后用另一部分对所建立的模型作检验，比较它们的预测误差，误差较小的模型和相应的变量即为我们所要选择的模型，这方面比较重要的方法是交叉核实 (Cross-Validation)。详见 Shao(1993)[8]、Raggi(2002)[9]、Rao 和 Wu(2005)[10] 等。此方法的主要缺陷是计算量大，而且依赖于所建立模型的方法。

三、信息论方法

这一类方法是基于 Kullback-Leibler 信息距离或最小化 Bayes 后验概率所建立的一套选择标准。事实上，它就是我们所熟知的 AIC 和 BIC 标准。详见 Akaike(1973)[11],Schwarz(1978)[12] 等。信息论方法在统计上来讲，需要借助模型的似然度，而似然函数一般情况需要知道分布的类型，仅有一些参数未知。对于总体分布完全未知的情形，可以借助非参数似然，也就是经验似然来进行，最近几年这种方法引起了一些学者的注意。详见 Variyath,Chen 和 Abraham(2010)[13]. 此方法较大的不足是对维数 p 较高时计算工作量较大。

四、惩罚方法

在变量选择中引入惩罚思想很早就引起了学者的重视。事实上，前三类方法有很大一部分都属于惩罚方法类。张军舰 (2008)[14] 详细地总结了惩罚思想在统计中的应用，此处简单介绍一下在作变量选择时惩罚思想的应用。如果记 $S_{RS}(p)$ 为模型 (1-1) 或 (1-2) 的残差平方和，即

$$S_{RS}(p)=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=Y^T(I-X(X^TX)^{-1}X^T)Y,$$

其中 $\hat{y}_i$ 为模型的拟合值。在变量选择时，一般情况下，希望 $S_{RS}(p)$ 愈小愈好，但仅按此标准，“最优”的模型是把所有的变量全部选进来，这是因为 $S_{RS}(p)$ 关于 p 是一个减函数。这显然是不合理的，必须对随 p 增加而减小的 $S_{RS}(p)$ 施加一个惩罚。经典统计中的许多变量选择标准均可归纳到惩罚思想中，例如：平均残差平方和 $\hat{\sigma}^2(p)$ 标准中，$\hat{\sigma}^2(p)=S_{RS}(p)/(n-p)$，惩罚因子 $1/(n-p)$；预测偏差的方差 $(n+p)\hat{\sigma}^2(p)=(n+p)S_{RS}(p)/(n-p)$，惩罚因子为 $(n+p)/(n-p)$；C_p 统计量 $C_p=S_{RS}(p)/\hat{\sigma}^2-n+2p$，其中 $\hat{\sigma}^2$ 为全模型的残差均方，与 p 无关，此时惩罚因子为 $2p$；$AIC=n\log(S_{RS}(p))+2p$，惩罚因子为 $2p$；$BIC=n\log(S_{RS}(p))+2p\log(n)$，惩罚因子为 $2p\log(n)$ 等等，详见王松桂等 [6] 或张军舰 (2008)[14]。

除前述经典的变量选择方法外，近年来许多学者借助回归系数估计方法来作变量选择，也就是把回归系数为 0 或近似为 0 的变量删除掉，这种方法需要先对回归系数作出估计. 但实际中有可能出现病态的设计阵，此时回归系数的估计就会极不稳定，出现较大的波动。这时惩罚思想的使用就会得到比较有效且稳健的估计和相应的变量选

择方法。例如岭估计、LASSO(Least absolute shrinkage and selection operator) 以及它们相应的推广等。详见 Breiman(1996)[15]、Tibshirani(1996)[16]、Fan 和 Li(2001)[17] 、Tibshirani 等 (2005)[18]、Zou 和 Hastic(2005)[19]、Turlach 等 (2005)[20]、Yuan 和 Lin(2006)[21]、 Zhao 和 Yu(2006)[22]、Wu 等 (2007)[23]、Meier 和 B$\ddot{u}$hlman(2007)[24]、Fan(2007)[25]、Zhang 和 Huang (2007)[26]、Hesterberg 等 (2008)[27]、Van.De.Geer.S.(2008)[28]、Zhao 等 (2009)[29]、Fan 和 Lv(2008,2009,2010)[30]−[32] 等等。

从上述描述我们可以看到，对解决这类问题，研究者提出了很多的惩罚思想，但是什么样的惩罚函数才是好的惩罚函数？Fan 和 Li(2001)[17] 指出，对于一个好的惩罚函数，它得出的解应该具有下面的三个方面的性质：无偏性 (Unbiasedness)、稀疏性 (Sparsity) 和连续性 (Continuity)。无偏性指的是，当真实未知参数 (即未知的回归系数) 不是零时，相应的估计应该是近似无偏的，这样可以减少不必要的模型偏差；稀疏性指的是回归系数的估计应该有一个限制，这个限制可以自动把一些较小 (接近于 0) 的估计系数去掉，并删除对应的变量，从而降低了模型的复杂性；连续性指的是所得的参数估计关于数据在某一范围内是连续的，从而避免在做预测时出现的不稳定现象。根据这些性质，Fan 和 Li(2001)[17] 构造了一个新的惩罚函数 $p_\lambda(\theta)$，其导函数 $p'_\lambda(\theta)$ 满足

$$p'_\lambda(\theta) = \lambda\{I(\theta \le \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda)\},$$

其中参数 $a > 2$, $\theta > 0$，该惩罚函数被称为绝对偏差的光滑剪切 (简称为 $SCAD$)。

在高维线性模型的变量选择中，比较突出的是两个问题：一是计算速度，二是统计的准确性。无论对于那种惩罚函数，我们最后都要依赖于一个好的算法，如果算法计算过于复杂，那么解决的难度还是很大，只有惩罚函数好，再加上一个好的算法才更为人们欢迎。例如使 $LASSO$ 方法广为流行的算法是由 Efron 等人 [33] 提出的“最小角回归算法”(Least Angle Regression, 简记为 LARS)。

前面这些方法对 $n < p$ 且 p 固定时，在某些准则下每种方法都有自己的优势，但当 $p \gg n$ 情形时，其计算难度还是比较大的，有时甚至可能无法计算。针对这样的问题，Candes 和 Tao(2007)[4] 提出了 Dantzig 选择方法。这种方法是在设计阵满足一致不确定原理 (Uniform uncertainty principle, 简记为 UUP) 条件下，利用 L_1 范数得到的一种变量选择方法，这种方法有一些较好的性质。例如：该问题的求解过程可以转化为一个线性规范问题，容易求解。但对超高维情况，这种方法的计算量较大，而且当维数增大时，UUP 条件不容易被满足，不能保证它能选择到正确模型。为解决 Dantzig 选择方法带来的问题，Fan(2007)[25] 和 Fan 和 Lv(2008)[30] 提出了安全独立筛选 (Sure independent screening, 简记为 SIS) 方法，该方法可以将维数 p 降到 d 使得 $d < n$，基于 SIS 方法，Fan 和 Lv(2008)[30] 给出一个非多项式的超高维 ($\log(p) = O(n^\xi), \xi > 0, p > n$) 线性模型的变

量选择方法。即先运用 SIS 方法把指标维数 p 降到 $d(d < n)$，然后再利用惩罚思想进行参数估计，最后得到相应的变量选择方法。不同的惩罚思想可以得到不同的变量选择方法，如：SIS-SCAD、SIS-Dantzig selector(DS)、SIS-Lasso、SIS-adaptive Lasso(AdaLasso)等。有时候也可以将维数降低后综合运用惩罚思想，如 SIS-DS-SCAD 或者 SIS-DS-AdaLasso 等，这些方法先采用降维然后再运用惩罚方法在计算速度等方面要比直接惩罚相对而言要快很多。同时针对高维问题 Fan 和 Lv(2008)[30] 还提出了先降维然后再采用分组方法来进行变量选择的思想。并简单介绍了两种方法：Subject related transformation 和 statistical transformation. 关于分组变量惩罚的详细内容可见 Yuan 和 Lin(2006)[21]，以及 Kim 等 (2006)[34]。在实际问题中，运用 SIS 方法时，有可能发生以下现象：一些不重要的变量可能由于与重要变量的共线性问题而被优先选取；还有一些变量单个来看可能与因变量相关程度不高，但其联合起来可能与因变量有较强的关系，此时运用 SIS 就选不上这些变量等。这些现象有可能降低 SIS 方法的正确率。因此 Fan 和 Lv(2008)[30] 还对 SIS 进行了扩展，提出了迭代 SIS(ISIS) 去克服这些问题。详见 Fan 和 Lv(2010)[32] 的综述文献。

由前述几个方面可以看出，尽管目前有关线性模型变量选择的方法非常多，但它们中一般都是基于模型误差为正态分布，并借助最小二乘或惩罚最小二乘思想或者更广的最小 L_q 方法或惩罚 L_q 方法 (详见 Wu 等 (2007)[23]) 建立。特别是对大 p 小 n，即维数 p 大于样本容量 n 的情况。前述方法在某些情况计算速度较慢，且准确性较差，有些时候由于惩罚函数的不同，得到的结果也不完全统一。这给应用也带来了较大的困难，尽管也有一些文章讨论一些改进策略，如 Peter(2006)[35] 的 Boosting 方法、Fan 和 Lv(2008)[30] 的 SIS、ISIS 方法等。但这些方法要么对误差项的要求是正态分布，要么仅讨论了相容性等基本性质。那么对误差分布未知时，如何能够合理充分地利用样本信息呢？如何才能得到它们较好的统计性质呢？我们认为，近年来在非参中比较流行的经验似然方法可以利用较多的样本信息。经验似然是 Owen(1988)[36] 系统提出和发展的一种非参数方法，它与已有的一些方法相比具有很多突出的优点。如基于经验似然构造的置信区间具有域保持性、变换不变性、置信域的形状由数据决定、Bartlett 纠偏性以及无需构造枢轴量等优点。详见王启华 (2004)[37] 的综述性文章。但直接用经验似然也存在有一些问题，除计算比较复杂外，它有一个前提条件，那就是由参数 θ 构造的估计方程 $E_F g(y,\theta) = 0$ 中，$\{g(y_i,\theta), i = 1,\cdots,n\}$ 的凸包必包含 0 点并且 0 为其一个内点。在实际应用和小样本中，这一点未必会满足，特别是小样本或者是参数真值与估计值有一定偏差时，经常会违背这一前提条件。Baggerly(1999)[38]、Owen(2001)[39]、Bartolucci(2007)[40]、Otsu(2007)[41]、Chen,Variyath 和 Abraham(2008)[42] 等均注意到这个问题，分别提出了经验似然 t 方法 (Studentized empirical likelihood)、惩罚经验似然 (Penalized empirical likelihood)、调整经验似然 (Adjusted empirical likelihood) 等方法。

针对前述各种方法的优劣，本文主要考虑误差项分布未知情况的高维线性模型的变量选择问题，计划将他们组合起来使用，使他们的优点得以发挥，同时又利用各自的优点弥补别的方法的一些劣势，尽可能找到一些要求条件不高、计算简单、准确性又高的变量选择方法。

§1.2 本文主要研究内容及成果

本文的主要研究思路和内容

本论文主要是研究如下情况的变量选择：指标之间的真实模型是线性模型，且维数 $s(< n)$ 较低，即满足稀疏性条件。但实际中的观测数据维数 $p \gg s$，甚至 $p \gg n$。如何从这 p 个指标中选出真正的 s 个指标且能够发现真实的模型。

本文的研究思路：首先借助 Fan(2007)[25] 和 Fan 和 Lv(2008)[30] 中所描述的 SIS 思想和算法，先选出与因变量相关性较强的 $m \triangleq [n\gamma]$ 个变量，把维数 p 降到 m(其中 $m < n$)；然后借助调整的经验似然 (AEL) 方法，对参数进行估计，并得到经验对数似然比函数，进而得到相应的 AIC 或 BIC 的经验似然版本；最后利用 AIC(或 BIC) 的经验似然版本进行具体的变量选择，得到最终的模型。这就是 SIS+AEL 方法，为提高变量选择的正确率，类似于 ISIS 思想，我们提出了由 SIS+AEL 方法通过迭代而得到的迭代 SIS+AEL 方法。即先用 SIS+AEL 方法从 $x_1, \cdots, x_p$ 中得到初步的变量 $x_{i1}, \cdots, x_{im_1}$，然后建立 y 与 $x_{i1}, \cdots, x_{im_1}$ 之间的线性模型，接着对其残差 $y - (x_{i1}\hat{\beta}_{i1} + \cdots + x_{im_1}\hat{\beta}_{im_1})$ 和剩余的 $p - m_1$ 个变量再应用 SIS+AEL 方法选择一部分变量，如此循环，直至得到我们满意的变量为止。

本论文的写作结构如下：全文共分五章：第一章绪论，介绍前人的研究成果和优劣比较，并给出本文的研究思路和研究内容；第二章介绍本文的主要工作，先分别介绍 Fan(2007)[25] 的 SIS 和 ISIS 方法，及 Chen,Variyath 和 Abraham(2008)[42] 的调整经验似然方法，然后介绍把二者融在一起的算法及有关问题；最后指出在计算过程中应注意的问题；第三章给出前述算法中的一些统计性质，从理论方面保证前述算法的可行性；第四章给出一些模拟结果，并与已有的一些结果进行对比；最后一章对全文进行总结并给出进一步研究的问题。

§1.3 本论文的主要成果和创新点

本论文的主要成果和创新之处主要体现在以下几点：

1. 对已有的一些方法进行重新组合，取长补短，降低了计算的工作量。我们的方法在普通计算机上很快就可以得到结果，例如 $p = 100, n = 20$，真实模型为 $s = 3$ 时，我们模拟一次所需的时间不到一秒钟。

2. 我们的方法去掉误差项是正态分布的限制，Fan(2007)[25] 的 SIS 和 ISIS 方法尽管简

单，但对误差项要求是正态分布，只有这样才能满足其相应的性质。本文从理论上说明误差项不必是正态分布，在较宽的条件下也可以得到 SIS 和 ISIS 方法相同的结论。对指标维数 p 降到样本容量 n 以下的情况，我们选择经验似然方法，无须对误差项作任何分布假定。

3. 采用调整经验似然方法作变量选择克服了经验似然的一些缺陷，众所周知，经验似然在使用时有一前提约束，即参数 θ 构造的估计方程 $E_F g(y,\theta) = 0$ 中，$\{g(y_i,\theta), i = 1,\cdots,n\}$ 的凸包必包含 0 点并且 0 为其一个内点。为避免此条件不满足时所引起的系统偏差，我们选用了 Chen,Variyath 和 Abraham(2008)[42] 的调整经验似然方法。

4. 模拟结果显示：在误差项服从正态分布情况下，我们的方法与 Fan 和 Lv(2008)[30] 的方法相差不大；在误差项不服从正态分布情况下，我们的方法也能得到较理想的结果，至少不比已有较为流行的 LASSO 方法差。

5. 本论文的结果容易推广。这主要有两个方面的含义：一是我们的结果计算代价不大，容易实现；二是限于篇幅问题，我们仅考虑了线性模型。事实上，对广义线性模型论文中所有结果均是成立的。

第二章　高维线性模型的变量选择

本章主要介绍我们所考虑的计算思想和算法。首先，我们介绍 SIS 和 ISIS 方法，然后介绍调整的经验似然方法，最后考虑如何把这二者组合在一起进行高维线性模型的变量选择以及相应的注意问题。

§2.1 SIS 方法和 ISIS 方法

现代科学技术和社会经济领域有大量复杂数据，复杂性之一是维数高，变量的个数可能高达数百甚至数千。此时，我们需要选出影响大的变量，建立简洁的模型。而传统的子集选择方法，由于可能的子集数太多，计算量太大而很难执行。针对线性模型，目前流行的是罚最小二乘，该方法可以同时进行选择变量和估计参数，从而大大提高了计算速度 (Fan 和 Li(2006)[43])。尽管利用罚最小二乘方法可以大大提高计算速度，但总体而言计算量还是较大，还需要花费较长的时间和精力，而且得到的估计不一定精确，因此针对维数 p 远大于样本容量 n 的高维情况，Fan 和 Lv(2008)[30] 提出了一种新的比较简单的降维方法－安全独立筛选 (SIS) 和迭代安全独立筛选 (ISIS) 方法。下面我们简单介绍 SIS 方法和 ISIS 方法。

令 $Y = (y_1, \cdots, y_n)^T$ 是 n 维独立响应变量，n 是样本容量。考虑线性回归模型 (1-1) 或 (1-2)

$$Y = X\beta + \varepsilon,$$

其中 $\beta = (\beta_1, \cdots, \beta_p)^T$ 是一个 p 维参数，$\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)^T, X = (X_1, \cdots, X_n)^T$ 是一个 $n \times p$ 的设计矩阵，为方便讨论，我们假定 X 为列标准化的矩阵，Y 为中心化向量。即 X 中每一列所代表的变量的样本均值为 0，样本标准差为 1，Y 的样本均值为 0。

令 $M_* \triangleq \{1 \leq i \leq p : \beta_i \neq 0\}$ 为我们感兴趣的真实稀疏模型的指标集，$s = |M_*|$ 代表 M_* 中元素的个数，也就是真实模型中回归系数不为 0 的个数。令 $w \triangleq (w_1, \cdots, w_p)^T = X^T Y$。对任意给定的 $\gamma \in (0, 1)$, 定义子模型

$$M_\gamma \triangleq \{1 \leq i \leq p : |w_i| \text{为前最大的} [n\gamma] \text{个}\},$$

其中 $[n\gamma]$ 表示 $n\gamma$ 的整数部分。

这样就可以把全模型指标集 $\{1, \cdots, p\}$ 降到一个子模型指标集 M_γ，其中元素的个数 $[n\gamma] < n$，这就是 Fan 和 Lv(2008)[30] 的 SIS 方法，其计算复杂性为 $O(np)$。

SIS 方法是一种截断式的选择方法，它首先是把 $w_i = \sum_{j=1}^{n} x_{ji} y_j$，按绝对值 $|w_i|$ 的大小

排序，然后强制选择前 $[n\gamma]$ 个最大的 $|w_i|$。Fan 和 Lv(2008)[30] 在一定条件下，证明了如下性质：当 $n \to \infty$ 时，对某一 $\gamma \in (0,1)$ 有

$$P(M_* \subset M_\gamma) \to 1. \tag{2-1}$$

这个结论对 $\gamma = \gamma_n \to 0$ 时，在一些合适的条件下也是成立的。详见 Fan 和 Lv(2008)[30] 第五节的定理 1。因此在某些约束条件下，我们可以把高维线性模型从 p 维降到 $[n\gamma] < n$ 维下，这样我们就可以利用一些传统的方法或惩罚思想进行变量选择。

在实际应用中，由于共线性、变量间的交互作用等因素的影响，单纯使用 SIS 方法可能错误率较高，特别是 n 较小，p 较大时。这是因为性质 (2-1) 是对 n, p 都充分大时，且 $\log(p) = O(n^\xi)$(ξ 为给定的正数) 成立，对小样本这个结论未必成立。为了使 SIS 方法在小样本和其假定条件不成立时仍然可以使用，Fan 和 Lv(2008)[30] 中提出了一种迭代方法－ISIS 方法，该方法作变量选择过程如下：

第一步：从全模型 $M \triangleq \{1, \cdots, p\}$ 中利用 SIS 与其它惩罚方法 (或已有的变量选择方法) 的组合方法选出一个子模型：$M^{(1)} = \{i_1, \cdots, i_{k_1}\}$；

第二步：建立 Y 与 $X_{i_1}, \cdots, X_{i_k}$ 之间的线性回归模型 $Y = X_{i_1}\beta_{i_1} + \cdots + X_{i_k}\beta_{i_k} + \varepsilon$；

第三步：在观测点上计算第二步所建模型的残差 $r_j (j = 1, \cdots, n)$, $r_j = y_j - X_{i_1}\hat{\beta}_{i_1} - \cdots - X_{i_k}\hat{\beta}_{i_k}$，其中 $\hat{\beta}_{i_1}, \cdots, \hat{\beta}_{i_k}$ 为相应方法所得的参数估计值；

第四步：用 r_i 替换 y_i，把 $M - M^{(1)}$ 看作全模型 M；

第五步：重复前述四步工作 l 次，可以得到指标集

$$A = \bigcup_{i=1}^{l} M^{(i)},$$

其中 $M^{(1)}, \cdots, M^{(l)}$ 为不相交的指标 $\{1, \cdots, p\}$ 的子集，A 中元素的个数小于 n。

应用前述 ISIS 方法，Fan 和 Lv(2008)[30] 作了一些模拟，模拟结果显示，这种方法可以大大提高变量选择的正确率，即使在样本容量 n 很小的情况，其包含真实模型的比率都接近于 1。

§2.2 变量选择的经验似然方法

似然是统计中非常基本的一种方法，在作变量选择时，也有许多学者采用似然方法，例如 Fan 和 Lv(2009)[31] 的惩罚似然方法、前已叙述的 AIC、BIC 等方法，但这些方法基本都属于参数似然方法，它们需要知道误差项的分布或分布类型。如果误差项的分布完全未知，则参数似然方法就会失效。那么在误差分布未知时，如何利用似然的思想和方法来作变量选择呢？经验似然将是一个重要的解决办法。

经验似然是由 Owen(1988)[36] 年系统提出和发展的一种非参数似然方法，在最近二十年内得到了许多统计学家、经济学家等学者的重视，其理论和应用工作得到了极大的发展，有关经验似然的详细研究情况可参见 Owen(2001)[39]、王启华(2004)[37]、Kitamura(2007)[44]、Chen 和 Keilegom(2009)[45] 等综述性文献。

有关线性模型的经验似然方法讨论的文献也比较多，如 Owen(1991)[46]、Chen 和 Keilegom(2009)[45] 等，具体见这些文献及其后的参考文献。因此可以借助经验似然方法来作线性模型的变量选择。

本小节我们主要讨论线性模型的经验似然方法变量选择，我们的思想和方法主要是参见 Chen,Variyath 和 Abraham(2008)[42]、Chen 和 Keilegom(2009)[45] 以及 Variyath,Chen 和 Abraham(2010)[13] 的文章。下面我们就来叙述相应的思想和计算思路。

§2.2.1 线性模型的经验似然估计

设所考虑的模型仍为 (1-1) 或 (1-2)，即

$$y_i = X_i^T\beta + \varepsilon_i, \qquad i = 1, \cdots, n.$$

假定 ε_i 相互独立，$E(\varepsilon_i|X_i) = 0, Var(\varepsilon_i|X_i) = \sigma^2(X_i) < \infty$，存在某一正数 b，使得 $|\varepsilon_i| \leq b, a.s.$，则 β 的最小二乘估计为下面估计方程的解。

$$\sum_{i=1}^{n} X_i(y_i - X_i^T\beta) = 0. \tag{2-2}$$

假设参数真值 β_0 为方程 $E(X_i(y_i - X_i^T\beta)|X_i) = 0$ 的唯一解。令 $p_i = F\{y_i\}, (i = 1, \cdots, n)$ 为数据的概率质量，则按照 Owen(1988)[36]、Owen(1991)[46] 的思想，有关 β 的经验似然 (EL) 为

$$L_n(\beta) = \max\left\{\prod_{i=1}^{n} p_i : p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i X_i(y_i - X_i^T\beta) = 0\right\}. \tag{2-3}$$

上式本质上为一个约束的拟似然 (profile likelihood)。利用 Lagrange 乘子法 (2-3) 可以转化为无约束的问题求解，其目标函数为

$$T = \sum_{i=1}^{n} \log(p_i) + \lambda_0(\sum_{i=1}^{n} p_i - 1) + \lambda^T \sum_{i=1}^{n} p_i X_i(y_i - X_i^T\beta). \tag{2-4}$$

分别对上式关于 p_i 求导并利用已有的约束关系，容易得到

$$\hat{p}_i = \frac{1}{n[1 + \lambda^T X_i(y_i - X_i^T\beta)]}, \quad i = 1, \cdots, n,$$

其中 λ 满足

$$\sum_{i=1}^{n}\frac{X_i(y_i-X_i^T\beta)}{1+\lambda^T X_i(y_i-X_i^T\beta)}=0. \tag{2-5}$$

上述求解过程也可以参见 Owen(1991)[46] 或 Qin 和 Lawless(1994)[47]，在后一篇文献中，$g(X_i,\beta)=X_i(y_i-X_i^T\beta)$，这样便可以得到

$$L_n(\beta)=\prod_{i=1}^{n}\frac{1}{n[1+\lambda^T X_i(y_i-X_i^T\beta)]},$$

其中 λ 满足 (2-5) 式，由此可得对数经验似然为

$$l_n(\beta)=\log\ L_n(\beta)=-\sum_{i=1}^{n}\log[1+\lambda^T X_i(y_i-X_i^T\beta)]-n\log\ n. \tag{2-6}$$

对给定的 β，(2-6) 中包含有参数 λ，这时的计算可利用 Owen(1990)[48] 的变换思想，把求 (2-3) 的最大值问题变换为求其对偶问题，即在约束条件

$$1+\lambda^T X_i(y_i-X_i^T\beta)\geq\frac{1}{n}\quad \text{对每一}\ i=1,\cdots,n \tag{2-7}$$

下最小化目标函数

$$Q(\lambda)=-\sum_{i=1}^{n}\log[1+\lambda^T X_i(y_i-X_i^T\beta)].$$

约束条件 (2-7) 来源于 $0\leq\ p_i\leq 1$ 的限制，$-Q'(\lambda)$ 刚好就是 (2-5) 的左边部分。

$$Q''(\lambda)=\sum_{i=1}^{n}\frac{X_i(y_i-X_i^T\beta)^2X_i^T}{[1+\lambda^T X_i(y_i-X_i^T\beta)]^2}$$

在 $D=\{\lambda:1+\lambda^T X_i(y_i-X_i^T\beta)\geq\frac{1}{n},1\leq\ i\leq\ n\}$ 上为正定阵。可以证明 D 为凸闭紧集。因此在 D 上 $Q(\lambda)$ 有唯一的最小值，故可以借助负梯度方向的牛顿迭代方法进行计算。

另一方面，众所周知，我们所考虑的似然函数 $L_n(\beta)$ 在无约束情况下，$\hat{p}_i=1/n,i=1,\cdots,n$ 时达到最大值 n^{-n}，此时对应的 $\lambda=0$。因此如果方程 (2-2) 存在有一个解 $\hat{\beta}$，则在 $\beta=\hat{\beta}$ 处 $L_n(\hat{\beta})=n^{-n}$。因此可以定义对数似然比

$$r_n(\beta)=-2\log(L_n(\beta)/L_n(\hat{\beta}))=2\sum_{i=1}^{n}\log[1+\lambda^T X_i(y_i-X_i^T\beta)]. \tag{2-8}$$

由 (2-8) 式，我们可以最小化 $r_n(\beta)$ 得到 β 的极大经验似然估计 $\tilde{\beta}$，即

$$\tilde{\beta}=\ \arg\min\ r_n(\beta),$$

进而可以得到 $\tilde{p_i}$ 和 $l_n(\beta)$。由于此处估计方程的个数与参数 β 的维数一致，所以 $\tilde{\beta}=\hat{\beta}$，即为方程 (2-2) 的解，如果估计方程个数大于参数个数，则可以应用 Qin 和

Lawless(1994)[47] 的方法计算。具体计算时可利用 Nelder 和 Mead(1965)[49] 的算法。

§2.2.2 调整的经验似然方法

在第一章我们曾说过，上一小节的经验似然方法在使用时有一个前提约束，那就是 $\{X_i(y_i - X_i^T\beta), i = 1, \cdots, n\}$ 的凸包要包含 0 点。这在实际中或者在小样本下有可能不满足，参见 Bartolucci(2007)[40]、Otsu(2007)[41]、Chen,Variyath 和 Abraham(2008)[42] 等文章的模拟和实例说明。本论文从小样本计算可靠性考虑，选用 Chen,Variyath 和 Abraham(2008)[42] 的方法。

令 $g_i = g(y_i;\beta) = X_i(y_i - X_i^T\beta), \bar{g}_n = \dfrac{1}{n}\sum\limits_{i=1}^{n} g_i$, 并定义 $g_{n+1} = -a_n\bar{g}_n$, 其中 a_n 为一与 n 有关正常量，于是调整的拟经验对数似然比函数定义为：

$$\begin{aligned} W^*(\beta) &= \inf\left\{-2\sum_{i=1}^{n+1}\log[(n+1)p_i] : p_i > 0, i = 1, \cdots, n+1; \sum_{i=1}^{n+1} p_i = 1; \sum_{i=1}^{n+1} p_i g(y_i;\beta) = 0\right\} \\ &= 2\sum_{i=1}^{n+1}\log\{1 + \lambda^T g(y_i;\beta)\}, \end{aligned} \tag{2-9}$$

其中 $\lambda = \lambda(\beta)$ 为

$$\sum_{i=1}^{n+1}\frac{g(y_i;\beta)}{1 + \lambda^T g(y_i;\beta)} = 0$$

的解。

关于 $W^*(\beta)$ 以及 $W^*(\tilde{\beta})$ 的求解方法可参见上一小节。Chen, Variyath 和 Abraham(2008)[42] 从理论上探讨了这种方法的渐近性质，并且给出了 a_n 的一些经验选取结果，如 $a_n = \max\{1, \log n/2\}$ 等。

§2.2.3 经验似然的信息论标准

对给定的 m 个自变量 $X_1, \cdots, X_m$ 和一个因变量 y，可以借用前两小节的方法计算 $W^*(\tilde{\beta})$。对于全模型指标集 $M = \{1, \cdots, p\}$，选取其中的 m 个指标共有 $t \triangleq C_p^m$ 种可能的组合，记为 $M_m^{(1)}, \cdots, M_m^{(t)}$。对每一种组合 $M_m^{(j)}$，均可以计算出相应的调整的拟经验对数似然值 $W^*(\tilde{\beta}(M_m^{(j)}))$, $j = 1, \cdots, t = C_p^m$。因此我们在 m 个指标所确定的 $\{W^*(\tilde{\beta}(M_m^{(j)}))$, $j = 1, \cdots, t = C_p^m\}$ 中，选择最小的一个作为 m 个指标中最优的模型集。并定义

$$W^*(m) = \min\left\{W^*(\tilde{\beta}(M_m^{(j)}))\text{, } j = 1, \cdots, t = C_p^m\right\},$$

进而我们可以定义 AIC 和 BIC 的调整经验似然版本

$$EAIC(m) = W^*(m) + 2m \tag{2-10}$$

$$EBIC(m) = W^*(m) + m\log(n) \tag{2-11}$$

于是针对不同的 m，我们可以得到不同的线性模型与相应的 EAIC 或 EBIC，EAIC 或 EBIC 达到最小时所对应的线性模型即为我们要寻找的最优模型，这种想法与 Variyath,Chen 和 Abraham[13] 是一致的。

§2.3 新的变量选择方法和算法

结合前述已有的研究，在本文中我们主要讨论的是当参数的维数 p 远大于样本容量 n 的情况，针对 $p \gg n$ 的情况，我们先采用 Fan 的 SIS 思想将维数降低到 m, 并且 $m < n$. 然后利用经验似然方法来进行变量选择，最终选出真实的模型。

具体算法步骤如下：

一、输入数据：

1、输入数据 Y, X，并把 X 标准化，Y 中心化，得数据阵如下：

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix} \triangleq \begin{pmatrix} X_1^T \\ \vdots \\ X_n^T \end{pmatrix} \triangleq (x_1, \cdots, x_p).$$

即 $\bar{Y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i = 0$, $\bar{X}_i = \dfrac{1}{n}\sum_{j=1}^{n} X_{ji} = 0$, $\dfrac{1}{n-1}\sum_{j=1}^{n} X_{ji}^2 = 1.$

二、SIS 方法 (初选)

1、$\sigma = 10^{-3}$, $m = n/\log n$;

2、计算 $w_i = X_i^T Y$ $(i = 1, \cdots, n)$ 或写成向量形式 $w = X^T Y$;

3、把 $|w_i|$ 按照从大到小排序，并选取其中的 m 个最大的 $|w_i|$，不妨记为 $|w|_{(1)}, \cdots, |w|_{(m)}$;

4、如果 $|w|_{(m)} > \sigma\sqrt{\frac{1}{n-1}\sum_{i=1}^{n} y_i^2}$，则继续下一步；否则 $m = m - 1$，继续判断，不妨记最后所选取的 w_i 为 $|w|_{(1)}, \cdots, |w|_{(m_1)}$, $m_1 \leq m$;

5、选取 $|w|_{(1)}, \cdots, |w|_{(m_1)}$ 所对应的自变量，不妨记其对应的观测分量为 $z_1, \cdots, z_{m_1}$，注意 $z_1, \cdots, z_{m_1}$ 为 $x_1, \cdots, x_p$ 的一个子集，其变量个数为 m_1。

三、调整经验似然变量选择

记 $Z=(z_1,\cdots,z_{m_1})_{n\times\ m_1}=\begin{pmatrix} Z_1^T \\ \vdots \\ Z_n^T \end{pmatrix}, Z_i^T=(Z_{i1},\cdots,Z_{im_1})$. 此时模型为

$$Y=Z\beta+\varepsilon,$$

此处 β 为 m_1 维向量，记 $\beta^{(0)}$ 为估计方程 $Z^T(Y-Z\beta)=0$ 的解，即

$$\beta^{(0)}=(Z^TZ)^{-1}Z^TY=\Big(\sum_{i=1}^{n}Z_iZ_i^T\Big)^{-1}\sum_{i=1}^{n}Z_iy_i. \tag{2-12}$$

另记

$$R(\lambda)=\sum_{i=1}^{n+1}\log(1+\lambda^Tg_i(\beta)),$$

$$R_\lambda=\sum_{i=1}^{n+1}\frac{g_i(\beta)}{1+\lambda^Tg_i(\beta)}\ ,\ R_{\lambda\lambda}=\sum_{i=1}^{n+1}-\frac{g_i(\beta)g_i^T(\beta)}{(1+\lambda^Tg_i(\beta))^2}.$$

1、赋初值：$\lambda_0=0,\gamma_1=1,\varepsilon_1=10^{-8}$;

2、按公式 (2-12) 计算 $\beta^{(0)}$ 的值;

3、计算 $g_i\triangleq\ g_i(\beta^{(0)})=Z_i(y_i-Z_i^T\beta^{(0)})\ ,i=1,\cdots,n\ ,\ \bar{g}_n=\dfrac{1}{n}\sum\limits_{i=1}^{n}g_i$;

4、给定 β 值，计算 λ 的值：

(1)、$k=0$;

(2)、计算 $R(\lambda_k),R_{\lambda_k},R_{\lambda_k\lambda_k},\Delta(\lambda_k)=-R_{\lambda_k\lambda_k}^{-1}R_{\lambda_k}$;

(3)、若 $\|\Delta(\lambda_k)\|<\varepsilon_1$, 转 (6)，否则继续下一步;

(4)、$\delta_k=\gamma_k\Delta(\lambda_k)$, 若对某一个 $i,1+(\lambda_k+\delta_k)^Tg_i(\beta)\le 1/n$ 或 $R(\lambda_k+\delta_k)<R(\lambda_k)$, 取 $\gamma_k=\gamma_k/2$ 重复 (4)，否则继续下一步;

(5)、$\lambda_{k+1}=\lambda_k+\delta_k,k=k+1,\gamma_{k+1}=(k+1)^{-1/2}$, 重复 (2)-(5);

(6)、计算 $W(\beta^{(0)})=\sum\limits_{i=1}^{n+1}\log\{1+\lambda^Tg_i(\beta^{(0)})\}$.

5、计算 $\tilde{\beta}$ 的值以及 $W^*(\tilde{\beta})$(其算法思想见 Nelder 和 Mead(1965)[49])

如果 g_i 的维数等于 $\beta^{(0)}$ 的维数，则 $\tilde{\beta}=\beta^{(0)}$，如果如果 g_i 的维数大于 $\beta^{(0)}$ 的维数，则按下述步骤计算：

(1)、初值的取法：$\alpha=\dfrac{1}{2},\dfrac{2}{3},1;\mu=2,3,4;\nu=\dfrac{1}{4},\dfrac{1}{3},\dfrac{1}{2}$;

(2)、在 $\beta^{(0)}$ 附近寻找 s 个点 $\beta^{(i)},\ i=1,2,\cdots,s$, 并且保证所选的这 s 个点满足：$1+\lambda^Tg_i(\beta^{(i)})>1/n$, 进而分别计算出 $W(\beta^{(i)})$;

(3)、定义：$\beta^{(h)}$ 为 $W_h=\max\{W(\beta^{(i)})\}$ 对应的点 , $\beta^{(l)}$ 为 $W_l=\min\{W(\beta^{(i)})\}$ 对应的点, $i=0,1,\cdots,s$. 选取 $\bar{\beta}=\frac{1}{s+1}\sum\limits_{i\neq\ h}\beta^{(i)},\ \beta^*=(1+\alpha)\bar{\beta}-\alpha\beta^{(h)},\ \alpha>0$. 计算 $W(\beta^*)$.

(4)、判断

(a) 若 $W(\beta^*) < W_l$, 则更新 $\beta^{**} = \mu\beta^* + (1-\mu)\bar{\beta}$, $\mu > 1$, 计算 $W(\beta^{**})$. 若 $W(\beta^{**}) < W_l$, 则 $W_h = W(\beta^{**})$; 若 $W(\beta^{**}) > W_l$, 则 $W_h = W(\beta^*)$.

(b) 若 $W(\beta^*) > W(\beta^{(i)})$, $i \neq h$, 则更新 $\beta^{**} = \nu\beta^{(h)} + (1-\nu)\bar{\beta}$, $0 < \nu < 1$, 计算 $W(\beta^{**})$. 若 $W(\beta^{**}) > W_h$, 则 $\beta^{(i)} = \frac{1}{2}(\beta^{(i)} + \beta^{(l)})$; 否则 $W_h = W(\beta^{**})$.

(5)、给定 $\kappa = 10^{-5}$, 当 $\sqrt{\frac{1}{s}\sum_{i=1}^{s+1}(W_i - \overline{W})^2} < \kappa$ 时, 其中 $\overline{W} = \frac{1}{s+1}\sum_{i=1}^{s+1} W_i$, 输出 $W^*(\tilde{\beta}) \triangleq \min\{W_i\}$ 及相应的 $\tilde{\beta}$，否则重复 (3)-(4);

6、得到 $\tilde{\beta}$ 和 $W^*(\tilde{\beta})$ 后，$\beta^{(0)} = \tilde{\beta}$，重复执行 3-5，如果 $|\tilde{\beta} - \beta^{(0)}| < 10^{-4}$，则进行下一步；

7、按照 (2-10) 和 (2-11) 计算 EAIC、EBIC；

8、取不同的 $t = 1, \cdots, m_1$，重复调整经验似然变量选择的全部过程，计算出相应的 EAIC、EBIC；

9、计算 minEAIC 和 minEBIC，便得到相应的设计变量 (或自变量) 和回归系数的估计。

四、ISIS 方法与调整经验似然方法的结合

由于采用 SIS+AEL 方法进变量选择包含真实模型的正确率较低，为提高包含真实模型的正确率，类似于 ISIS 思想，我们提出了迭代 SIS+AEL 方法。具体算法如下：

第一步：从全模型 $M \triangleq \{1, \cdots, p\}$ 中利用 SIS 与 AEL 的组合方法选出一个子模型：$M^{(1)} = \{i_1, \cdots, i_{k_1}\}$；

第二步：建立 Y 与 $X_{i_1}, \cdots, X_{i_k}$ 之间的线性回归模型 $Y = X_{i_1}\beta_{i_1} + \cdots + X_{i_k}\beta_{i_k} + \varepsilon$；

第三步：在观测点上计算第二步所建模型的残差 $r_j (j = 1, \cdots, n)$, $r_j = y_j - X_{i_1}\hat{\beta}_{i_1} - \cdots - X_{i_k}\hat{\beta}_{i_k}$，其中 $\hat{\beta}_{i_1}, \cdots, \hat{\beta}_{i_k}$ 为相应方法所得的参数估计值；

第四步：用 r_i 替换 y_i，把 $M - M^{(1)}$ 看作全模型 M；

第五步：重复前述四步工作 l 次，可以得到指标集

$$A = \bigcup_{i=1}^{l} M^{(i)},$$

其中 $M^{(1)}, \cdots, M^{(l)}$ 为不相交的指标 $\{1, \cdots, p\}$ 的子集，A 中元素的个数小于 n。

§2.4 新的变量选择算法中应注意的问题

§2.3 的算法尽管看起来比较复杂，但真实计算并不是很复杂，因为向量或矩阵运算在现有的统计软件如 R、Matlab 中都有现成的语句。在真实应用中，应该注意以下几点：

1、运用 (2-12) 计算初值 $\beta^{(0)}$ 时，如果 $\sum_{i=1}^{n} Z_i Z_i^T$ 出现不可逆时，此时可适当降低变量

的维数 m_1. 换句话说，此时可按照 SIS 选择 m_1-1 个变量，如果还不满足 $\sum_{i=1}^{n} Z_i Z_i^T$ 可逆条件，则继续降低维数，选 m_1-2，如此进行，直至出现相对应的 $\sum_{i=1}^{n} Z_i Z_i^T$ 可逆时停止。

2、在第 5 步计算 $\tilde{\beta}$ 的值时，我们的估计方程 g_i 的维数与参数维数是相同的。故并没有用到 g_i 的维数大于参数维数的情况，但在实际计算时，有可能出现估计方程的维数大于参数维数情况，作为一个完整的算法还是写上比较完整些。此处没有应用牛顿迭代是因为在 $\tilde{\beta}$ 处无法从理论上证明它有凸凹性，所以只能用 Nelder 和 Mead(1965)[49] 的单纯形方法。

第三章　新的变量选择方法的性质

结合 SIS 方法与调整的经验似然方法来进行变量选择仍然保留着直接利用 SIS 方法降维及调整的经验似然方法所拥有的相应的一些结论与性质。

由模型 (1-2) 可以有

$$Y = X\beta + \varepsilon.$$

记 $M_* = \{1 \leq\ i \leq\ p : \beta_i \neq 0\}$ 为我们感兴趣的真实稀疏模型，$s = |M_*|$ 为 M_* 中的元素的个数。假定观测设计阵 X 为一个 p 维总体 x 的 n 次观测所形成的矩阵，总体 x 的均值为 0，协方差阵为 Σ，Σ 的对角线上元素全为 1。定义：

$$z = \Sigma^{-1/2}x\ \ ,\ \ Z = X\Sigma^{-1/2}.$$

由此知设计阵 Z 为独立同分布的变量。

为方便后面讨论，我们先给出如下条件和引理：

条件 1: $p > n$ 且对某个 $\xi > 0$, 有 $\log(p) = O(n^{\xi})$.

条件 2: z 满足球对称分布及性质 C.

性质 C 即同心性质 (Concentration Property) 的含义 (见 Fan 和 Lv(2008)[30])：存在某个 $c, c_1 > 1$ 及 $C_1 > 0$，使得下面不等式

$$P(\lambda_{\max}(\tilde{p}^{-1}\tilde{Z}\tilde{Z}^T) > c_1 \text{及} \lambda_{\min}(\tilde{p}^{-1}\tilde{Z}\tilde{Z}^T) < 1/c_1) \leq\ e^{-C_1 n}$$

成立，则我们说矩阵 Z 具有同心性质。其中 $\tilde{Z}$ 为 Z 的一个 $n \times \tilde{p}$ 的 Z 的子矩阵且 $cn < \tilde{p} \leq\ p$. 其中 $\lambda_{\max}(\cdot)$ 与 $\lambda_{\min}(\cdot)$ 分别表示矩阵的最大与最小特征根。

条件 3: $\mathrm{var(Y)} = \mathrm{O(1)}$ 及对某个 $\kappa \leq 0, c_2, c_3 > 0$, 有

$$\min_{i \in\ M_*} |\beta_i| \geq \frac{c_2}{n^{\kappa}}, \text{及}\ \ \min_{i \in\ M_*} |cov(\beta_i^{-1}Y, X_i)| \geq\ c_3.$$

条件 4: 存在 $\tau \geq 0$ 及 $c_4 > 0$ 使得

$$\lambda_{\max}(\Sigma) \leq\ c_4 n^{\tau}.$$

条件 5: 假定 ε_i 相互独立，$E(\varepsilon_i|X_i) = 0\ , 0 < Var(\varepsilon_i|X_i) = \sigma^2(X_i)$，存在某一正常数 b_i，使得 $|\varepsilon_i| \leq\ b_i\ \ a.s.$。

条件 5 作为误差项的要求比 Fan 和 Lv(2008)[30] 的正态要求减弱了许多，而且在实际数据应用中也比较容易满足。事实上，如果去掉界值 b_i 的限制也是成立的。详见下面的

引理 3.1 说明。

性质 C 实际上是对设计阵 Z 的一个约束。直觉上，它说明 $n \times \tilde{p}$ 阶阵 $\tilde{Z}$ 的奇异值以较大的概率是同阶的。事实上，由于当 $\tilde{p} \to \infty$ 时，$\tilde{p}^{-1}\tilde{Z}\tilde{Z}^T \to I_n$，所以当 $\tilde{p}$ 越大，$\tilde{p}^{-1}\tilde{Z}\tilde{Z}^T$ 越接近于 I_n，故性质 C 感觉是很自然的，详细的理论说明可参见 bai 和 Silverstein(2006)[50]，其它条件的合理性解释见 Fan 和 Lv(2008)[30]。

引理 3.1 设 $\xi_1, \cdots, \xi_n$ 是 n 个独立的随机变量，$E\xi_i = 0$, $\sigma_i^2 = E\xi_i^2$, $|\xi_i| \leq b_i \ a.s.(\forall\, i \geq 1)$, $b = \max\limits_{1 \leq j \leq n} b_j$, $\sigma^2 = \dfrac{1}{n}\sum\limits_{i=1}^{n}\sigma_i^2$，则对 $\forall\, \varepsilon > 0$,

$$P\Big(|\frac{1}{n}\sum_{i=1}^{n}\xi_i| \geq \varepsilon\Big) \leq \exp\Big(-\frac{n^2\varepsilon^2}{2bn\varepsilon + 4n\sigma^2}\Big) = 2e^{-Cn},$$

其中 $C = \dfrac{\varepsilon^2}{2b\varepsilon + 4b^2}$.

引理 3.1 实际上就是 Bernstein 不等式，其证明比较经典，详见一些不等式教材或一些文献，如：Bernett(1962)[51] 等。引理 3.1 的有界性条件也可以去掉，详见张涤新、王志诚(2001)[52]，此处不再细述。

记

$$M_\gamma \triangleq \{1 \leq\ i \leq\ p : |w_i| \text{为前最大的}[n\gamma]\text{个}\},$$

其中 $w \triangleq (w_1, \cdots, w_p)^T = X^TY$, $\gamma \in (0, 1)$, $[n\gamma]$ 表示 $\gamma\, n$ 的整数部分。

定理 3.1　在上述条件 1-5 下，如 $2\kappa + \tau < 1$, 存在某个 $\theta < 1 - 2\kappa - \tau$, 使得当 $\gamma \sim\ cn^{-\theta}$ 且 $c > 0$ 时，存在 $C > 0$, 有

$$P(M_* \subset\ M_\gamma) = 1 - O(\exp(-Cn^{1-2\kappa}/\log\ n)).$$

证明 完全类似于 Fan 和 Lv[30]，只须把他们证明过程中应用 Lemma 3 的地方换为引理 3.1 即可。由于证明过程比较繁琐，此处略去具体细节。

由此可知，对充分大的样本容量 n，选择的变量包含真实变量的概率接近于 1。

下面我们考虑调整经验似然的性质。调整的经验似然与经验似然类似，也具有相同的渐近性质，其中比较重要的一个结果就是经验似然比在真值处的极限分布。

定理 3.2　设 $y_1, y_2, \cdots, y_n\ iid$ 于 q 维未知分布 F，β_0 为参数真值，满足 $E\{g(Y, \beta) : F\} = 0$. $g(Y, \beta) = X^T(Y - X\beta)$ 为一个 m 维函数。假设 $Var\{g(Y, \beta) : F\}$ 有限且 $m > q$. 对调整的拟经验对数似然比函数

$$W^*(\beta) = \sup\left\{\sum_{i=1}^{n+1}\log[(n+1)p_i] : p_i \geq 0, i = 1, \cdots, n+1; \sum_{i=1}^{n+1}p_i = 1; \sum_{i=1}^{n+1}p_i g_i = 0\right\}$$

及 $a_n = o_p(n^{2/3})$, 当 $n \to \infty$ 时, 有

$$-2W^*(\beta_0) \to \chi^2_m.$$

定理 3.3　假定定理 3.2 的条件满足，且对 $\beta \neq \beta_0$, 若 $\left\|E\{X^T(Y - X\beta)\}\right\| > 0$ 则有

$$-2n^{-1/3}W^*(\beta) \to \infty及 - 2n^{-1/3}W(\beta) \to \infty,$$

其中 $W(\beta) = \sum_{i=1}^{n} \log(n\hat{p}_i) = -\sum_{i=1}^{n} \log\{1 + \lambda^T g(y_i, \beta)\}$.

定理 3.4　假定定理 3.2 的条件满足，且 $\|g(Y, \beta)\|^3 = \left\|X^T(Y - X\beta)\right\|^3$ 在 β_0 的某个邻域内有界。$|\hat{\beta} - \beta_0| < n^{-1/3}$ 且 $\hat{\beta}$ 为 $W^*(\beta)$ 的一个局部最大值点，则当 $n \to \infty$ 时，有

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, \Sigma),$$

其中 $\Sigma = \{E(\partial\, g/\partial\beta)^T(Egg^T)^{-1}E(\partial\, g/\partial\beta)\}^{-1}$.

定理 3.2-3.4 的证明与经典的经验似然过程证明类似且与 Chen,Variyath 和 Abraham(2008)[42] 几乎一致，详见 Chen,Variyath 和 Abraham(2008)[42]，此处略去细节。

第四章　模拟研究

本章主要采用第二章的算法进行模拟研究，具体模型及模拟得到的相关结果如下：

一、模拟 I

1. 模拟条件

(1) 真实模型：$y = 5x_1 + 5x_2 + 5x_3 + \varepsilon.$

(2) 误差项 ε 的分布：分别取 $\varepsilon \sim N(0,1)$, $\varepsilon \sim U(-1,1)$, $\varepsilon \sim CN(0,1)$ 三种情况。其中 $N(\mu,\Sigma)$ 表示均值为 μ，方差为 Σ 的正态分布，$U(-1,1)$ 表示 $(-1,1)$ 上的均匀分布，$CN(0,1)$ 表示标准正态分布 $N(0,1)$ 在 $[-2,2]$ 上的截断分布，即

$$X \sim N(0,1)\ ,\ \varepsilon = \begin{cases} X, & X \in (-2,2), \\ -2, & X \leq -2, \\ 2, & X \geq 2 \end{cases}$$

(3) 设计阵要求：$(x_1,\cdots,x_p) \sim N(0,\Sigma)$,其中 $\Sigma = (\sigma_{ij})_{p\times p}$, $\sigma_{ii} = 1,(i = 1,\cdots,p),\sigma_{ij} = \rho,(i \neq j)$；$p$ 分别取 100,1000 两种情况；ρ 分别取 0,0.1,0.5,0.9 等几种情况。

(4) 样本容量分别选取 $n = 20, n = 50$。

(5) 模拟重复次数：200 次。

2. 模拟结果：

针对模拟条件中的各种搭配，我们计算了已有的 SIS、ISIS、LASSO、新方法 SIS+AEL 和迭代 SIS+AEL(I+(SIS+AEL)) 所包含真实变量的比率。所得模拟的具体结果如表 1 －表 3 所示：

表 1　当误差项 $\varepsilon \sim N(0,1)$ 时

p	n	方法	$\rho = 0$	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.9$
100	20	SIS	0.755	0.855	0.690	0.670
		ISIS	1.000	1.000	1.000	1.000
		LASSO	0.970	0.990	0.985	0.870
		SIS+AEL	0.355	0.440	0.200	0.230
		I+(SIS+AEL)	0.920	0.875	0.850	0.835

p	n	方法	$\rho=0$	$\rho=0.1$	$\rho=0.5$	$\rho=0.9$
100	50	SIS	1.000	1.000	1.000	1.000
		ISIS	1.000	1.000	1.000	1.000
		LASSO	1.000	1.000	1.000	1.000
		SIS+AEL	0.950	0.935	0.920	0.915
		I+(SIS+AEL)	1.000	1.000	1.000	1.000
1000	20	SIS	0.205	0.255	0.145	0.085
		ISIS	1.000	1.000	1.000	1.000
		LASSO	0.340	0.555	0.556	0.220
		SIS+AEL	0.060	0.080	0.040	0.020
		I+(SIS+AEL)	0.385	0.420	0.400	0.210
1000	50	SIS	0.990	0.960	0.870	0.860
		ISIS	1.000	1.000	1.000	1.000
		LASSO	1.000	1.000	1.000	1.000
		SIS+AEL	0.950	0.935	0.750	0.700
		I+(SIS+AEL)	0.965	0.930	0.915	0.920

表 2　当误差项 $\varepsilon \sim U(-1,1)$ 时

p	n	方法	$\rho=0$	$\rho=0.1$	$\rho=0.5$	$\rho=0.9$
100	20	SIS	0.655	0.755	0.650	0.620
		ISIS	1.000	1.000	1.000	1.000
		LASSO	0.870	0.830	0.840	0.810
		SIS+AEL	0.440	0.420	0.290	0.260
		I+(SIS+AEL)	0.880	0.845	0.850	0.835
100	50	SIS	0.975	0.935	0.950	0.925
		ISIS	1.000	1.000	1.000	1.000
		LASSO	0.935	0.940	0.920	0.915
		SIS+AEL	0.900	0.915	0.895	0.880
		I+(SIS+AEL)	0.940	0.945	0.925	0.915
1000	20	SIS	0.200	0.215	0.125	0.065
		ISIS	0.985	0.960	0.910	0.895
		LASSO	0.285	0.410	0.395	0.200
		SIS+AEL	0.030	0.070	0.020	0.030
		I+(SIS+AEL)	0.325	0.420	0.430	0.215

p	n	方法	$\rho=0$	$\rho=0.1$	$\rho=0.5$	$\rho=0.9$
1000	50	SIS	0.960	0.935	0.825	0.820
		ISIS	0.995	0.980	0.935	0.920
		LASSO	0.915	0.920	0.900	0.895
		SIS+AEL	0.925	0.930	0.895	0.885
		I+(SIS+AEL)	0.950	0.935	0.920	0.915

表 3　当误差项 $\varepsilon \sim CN(0,1)$ 时

p	n	方法	$\rho=0$	$\rho=0.1$	$\rho=0.5$	$\rho=0.9$
100	20	SIS	0.765	0.850	0.675	0.650
		ISIS	1.000	1.000	1.000	1.000
		LASSO	0.975	0.985	0.975	0.845
		SIS+AEL	0.390	0.410	0.310	0.250
		I+(SIS+AEL)	0.900	0.855	0.835	0.820
100	50	SIS	1.000	1.000	1.000	1.000
		ISIS	1.000	1.000	1.000	1.000
		LASSO	1.000	1.000	1.000	1.000
		SIS+AEL	0.935	0.920	0.900	0.910
		I+(SIS+AEL)	1.000	1.000	1.000	1.000
1000	20	SIS	0.195	0.220	0.125	0.070
		ISIS	1.000	1.000	1.000	1.000
		LASSO	0.345	0.565	0.536	0.205
		SIS+AEL	0.060	0.060	0.020	0.010
		I+(SIS+AEL)	0.375	0.410	0.435	0.210
1000	50	SIS	0.985	0.945	0.860	0.845
		ISIS	1.000	1.000	1.000	1.000
		LASSO	1.000	1.000	1.000	1.000
		SIS+AEL	0.925	0.910	0.735	0.695
		I+(SIS+AEL)	0.945	0.935	0.900	0.905

3. 结果分析

由表 1 －表 3 可以得到如下结论：

(1) 若 $p \gg n$，且 p 固定，n 增加时，包含正确模型的比率增加；

(2) 若 $p \gg n$，且 n 固定，p 增加时，包含正确模型的比率降低；

(3) 当变量 $x_1,\cdots,x_p$ 中两两相关性程度越高 (ρ 越大)，包含正确模型的比率越低；

(4) 在误差项服从正态分布时，包含正确模型的比率与 LASSO(现阶段比较流行的方法) 相差不大；但对误差项不服从正态分布时，整体上稍好于 LASSO。

二、模拟 II

1. 模拟条件

(1) 真实模型：$y=5x_1+5x_2+5x_3-15\sqrt{\rho}x_4+\varepsilon$；

(2) 误差 ε 的分布：分别取 $\varepsilon\sim\ N(0,1)$, $\varepsilon\sim\ U(-1,1)$, $\varepsilon\sim\ CN(0,1)$ 三种情况。

(3) 设计阵要求：$(x_1,x_2,x_3,x_5,\cdots,x_p)\sim\ N(0,\Sigma)$,其中 $\Sigma=(\sigma_{ij})_{p\times\ p}$, $\sigma_{ii}=1,(i=1,\cdots,p),\sigma_{ij}=\rho,(i\neq\ j)$；$x_4\sim\ N(0,1)$，并且与其余的 $p-1$ 个变量的相关性为 $\sqrt{\rho}$，p 分别取 100,1000 两种情况；ρ 取 0.5。

(4) 样本容量分别选取 $n=20,n=50,n=70$。

(5) 模拟重复次数：200 次。

2. 模拟结果：

针对模拟条件中的各种搭配，我们计算了已有的 SIS、ISIS、LASSO、新方法 SIS+AEL 和迭代 SIS+AEL(I+(SIS+AEL)) 所包含真实变量的比率。所得模拟的具体结果如表 4 －表 6 所示：

表 4　当误差项 $\varepsilon\sim\ N(0,1)$ 时

p	$\rho=0.5$ 方法	$n=20$	$n=50$	$n=70$
100	SIS	0.025	0.490	0.740
	ISIS	1.000	1.000	1.000
	LASSO	0.000	0.360	0.915
	SIS+AEL	0.000	0.400	0.570
	I+(SIS+AEL)	0.560	0.720	0.865
1000	SIS	0.000	0.000	0.000
	ISIS	1.000	1.000	1.000
	LASSO	0.000	0.000	0.000
	SIS+AEL	0.000	0.000	0.000
	I+(SIS+AEL)	0.325	0.415	0.530

表 5　当误差项 $\varepsilon \sim U(-1,1)$ 时

p	$\rho = 0.5$	方法	$n = 20$	$n = 50$	$n = 70$
100		SIS	0.020	0.435	0.650
		ISIS	1.000	1.000	1.000
		LASSO	0.000	0.280	0.735
		SIS+AEL	0.000	0.385	0.685
		I+(SIS+AEL)	0.525	0.715	0.830
1000		SIS	0.000	0.000	0.000
		ISIS	1.000	1.000	1.000
		LASSO	0.000	0.000	0.000
		SIS+AEL	0.000	0.000	0.000
		I+(SIS+AEL)	0.375	0.450	0.565

表 6　当误差项 $\varepsilon \sim CN(0,1)$ 时

p	$\rho = 0.5$	方法	$n = 20$	$n = 50$	$n = 70$
100		SIS	0.030	0.495	0.730
		ISIS	1.000	1.000	1.000
		LASSO	0.000	0.375	0.910
		SIS+AEL	0.000	0.395	0.555
		I+(SIS+AEL)	0.545	0.745	0.885
1000		SIS	0.000	0.000	0.000
		ISIS	1.000	1.000	1.000
		LASSO	0.000	0.000	0.000
		SIS+AEL	0.000	0.000	0.000
		I+(SIS+AEL)	0.320	0.425	0.555

3. 结果分析

模拟 II 与模拟 I 主要区别在于模拟 II 中的真实模型的自变量 x_4 与其余 $p-1$ 个变量的相关性不同。尽管如此，通过模拟，我们可以看到所得的模拟结果还是与模拟 I 结果类似，即若 $p \gg n$，且 p 固定，n 增加时，包含正确模型的比率增加；若 $p \gg n$，且 n 固定，p 增加时，包含正确模型的比率降低；在误差项服从正态分布时，包含正确模型的比率与 LASSO(现阶段比较流行的方法) 相差不大；但对误差项不服从正态分布时，整体上稍好于 LASSO。

第五章　总结和进一步问题

本论文在对前人已有研究成果进行综合分析的基础上，主要考察随机误差项分布未知时的高维线性模型的变量选择问题。所采用的方法和思想主要是 Fan 和 Lv 的 SIS 方法和 ISIS 方法，与 Chen,Variyath 和 Abraham(2008) 的调整经验似然方法的有机结合。从理论上进一步说明 Fan 和 Lv(2008) 的 SIS 方法和 ISIS 方法在误差项不服从正态分布时也是成立的，并且具有较好的安全筛选渐近性质。从算法上分别给出了 SIS 方法、迭代 SIS 方法与调整经验似然有机结合的 SIS+AEL 算法与迭代 SIS+AEL 算法。这个算法既保留了原有方法的渐近性质，又降低了实际中对误差项的分布要求，取长补短、计算简单、想法直观。模拟结果表明，我们的算法在对高维线性模型作变量选择时，若误差项服从的是正态分布，其包含真实模型的比率与 LASSO 相差不大，而当误差项不服从正态分布时，其结果整体上要略好于 LASSO。

本文可以进一步研究的问题有：

1. 经验似然计算相对复杂。Owen(1988)、Qin 和 Lawless(1994) 的论文中都指出，经验欧氏似然与经验似然具有相同的渐近性质，但经验欧氏似然计算相对简单。所以此处若将经验似然中的似然距离 $\log \prod_{i=1}^{n} np_i$ 改为欧氏距离 $-\frac{1}{2}\sum_{i=1}^{n}(p_i - 1/n)^2$，即采用经验欧氏似然代替经验似然来进行变量选择，则能进一步提高计算的速度。

2. 本文所讨论的线性模型结论很容易推广到广义线性模型，限于篇幅和精力，此处不再讨论。

3. 本文所讨论的方法和计算可以用来处理金融市场中的投资组合问题以及进行医学数据 (如：白血病数据) 分析，我们下一步将选择一些实际数据进行实证分析。

参 考 文 献

[1] T.J.Hastie,R.Tibshirani,J.Friedman.The elements of statistical learning:Data mining inference and prediction[J].Springer.2001.

[2] B.A.Turlach.On algorithms for solving least squares problems under an L_1 Penalty or an l_1 Constraint[J]. http://www.stat.nus.edu.sg/ statba/psfiles/jsm2004.pdf.

[3] B.Peter.Variable selection for high-dimensional data:with applications in molecular biology[J]. ftp://ftp.stat.math.ethz.ch/U/maechler/Manuscripts/buhlmann/ISI2007.pdf.

[4] E.Candes,T.Tao.The Dantzig selector:statistical estimation when p is much large than n (with discussion)[J].Annals of Statistics.2007,35:2313-2351.

[5] 王大荣. 分散度量模型中的变量选择 [D]. 北京：北京工业大学博士学位论文.2009.

[6] 王松桂, 陈敏，陈立萍. 线性统计模型 [M]. 北京：高等教育出版社.2000.

[7] A.J.Miller.Subset selection in regression[M].London:Chapman and hall,2002.

[8] J.Shao.Linear model selection by cross-validation[J]. Journal of the American Statistical Association.1993,88:486-494.

[9] M.Raggi. Asymptotic properties of cross-validation methods in regression problems[J]. PHD.thesis.Dipartimento discienze statistical university di bologna.2002.

[10] C.R.Rao and Y.Wu.Linear model selection by cross-validation[J].Journal of the American Statistical Planning and Inference.2005,128(1):132-240.

[11] H.Akaike.Information Theory and an extension of the maximum likelihood principle[J].B N Petrov,F Csaki,eds.Proceedings of the second international symposium on information theory.budapest,1973.267-281.

[12] G.Schwarz.Estimating the dimensions of a model[J].Annals of Statistics.1978,6:461-464.

[13] A.M.Variyath,J.Chen,B. Abraham.Empirical likelihood based variable selection[J].Journal of Statistical Planning and Inference 140(2010)971-981.

[14] 张军舰. 统计中的惩罚思想 [J].2008.(手稿).

[15] L.Breiman.Heuristics of instability and stabilization in model selection[J].Annals of Statistics.1996,24:2350-2383.

[16] R.Tibshirani.Regression shrinkage and selection via the Lasso[J].Journal of the Royal Statistical Society Series B.1996,58:267-288.

[17] J.Fan,R.Li.Variable selection via nonconcave penalized likelihood and its oracle properties[J].Journal of the American Statistical Association.2001,96:1348-1360.

[18] R.Tibshirani,S.Rosset,J.Zhu,K.Knight.Sparsity and smoothness via the fused lasso[J].Journal of the Royal Statistical Society Series B.2005,67,91-108.

[19] H.Zou,T.Hastie.Regularization and variable selection via the elastic net[J].Journal of the Royal Statistical Society Series B.2005,67:301-320.

[20] B.A.Turlach,W.N.Venables and S.J.Wright. Simultaneous variable selection[J]. Technometrics .2005,47:349-363.

[21] M.Yuan,Y.Lin.On the non-negative garrotte estimator[J].Journal of the Royal Statistical Society Series B.2007,69:143-161.

[22] P.Zhao, B.Yu. On model selection consistency of lasso [J]. J. Machine Learning reserch. 2006,7,2541-2567.

[23] T.Wu,Y.Chu,Y.Yu.On solving l_q-penalized regressions[J].Journal of Applied Mathematics and Decision Sciences.2007,Article ID 24053,13 pages.

[24] L.Meier,P.B$\ddot{u}$hlmann.Smoothing L_1-penalized estimators for high-dimensional time-course data[J].Electronic Journal of Statistics.2007,1:597-615.

[25] J.Fan.,Variable screening in high-dimensional feature space,ICCM 2007.VOL.II 735-747.

[26] C.H.Zhang,J.Huang.The sparsity and bias of the LASSO selection in high-dimensional linear regression[J].Annals of Statistics.2008,36(4):1567-1594.

[27] T.Hesterberg,N.H.Choi,L.Meier,C.Fraley.Least angle and l_1 penalized regression:A review[J].Statistics Surveys.2008,61-93.

[28] VAN.DE.GEER.S.High-dimensional generalized linear models and the LASSO[J].Annals of Statistics.2008,36(2):614-645.

[29] P.Zhao,G.Rocha,B.Yu.Grouped and hierarchical model selection through composite absolute penalties[J].Annals of Statistics.2009,37:3468-3497.

[30] J.Fan,J.Lv.Sure independence screening for ultra-high dimensional feature space[J]. Journal of the Royal Statistical Society Series B.2008,70:849-911.

[31] J.Fan,J.Lv. Non-concave penalized likelihood with NP-dimensionality[J]. 2009. http:// arxiv.org / PS_{cache} /arxiv/ pdf / 0910/ 0910.1119v1.pdf.

[32] J.Fan,J.Lv. A selective overview of variable selection in high dimensional feature space[J]. Statistica Sinica.2010,20:101-148.

[33] B.Efron,T.Hastie,I.ohnstone,R.Tibshirani.Least angle regression(with discussions)[J].Annals of Statistics.2004,32:409-499.

[34] Y,Kim.J,Kim.Y,Kim.Blockwise sparse regression[J].Statistic Sinnica.2006,16:375-390.

[35] B.Peter.Boosting for high-dimensional linear models[J].Annals of Statistics.2006,34(2):559-

583.

[36] A.B.Owen.Empirical likelihood ratio confidence interval for a single function [J]. Biometrika,1988,75:237-249.

[37] 王启华. 经验似然统计推断方法发展综述 [J]. 数学进展.2004,33(2):141-151.

[38] K.A.Baggerly.Studentized empirical likelihood and maximum entropy[J]. Technical Report,Department of Statistics,Rice University.1999.

[39] A.B.Owen.Empirical likelihood[J].London.Chapman&hall.2001.

[40] F.Bartolucci.A penalized version of the empirical likelihood ratio for the population mean[J]. Statistical & Probability letters.2007,77:104-110.

[41] T.Otsu.Penalized empirical likelihood estimation of semiparametric models[J]. J Multivariate. Analysis.2007,98:1923-1954.

[42] J.Chen,A.M.Variyath,B. Abraham.Adjusted empirical likelihood and its properties[J].Journal of Computational and Craphical Statistics.2008,17:426-443.

[43] J.Fan,R.Li.,Statistical challenges with high dimensionality:feature selection in knowledge discovery[J].Proceedings of the international congress of mathematicians(M.Sanz-Sole,J.Soria,J.L.Varona,J.Verdera,eds.)2006,Vol.III,595-622.

[44] Y.Kitamura.Empirical likelihood methods in econometrics: theory and practice[J]. Econoometric society monographs. 2007,43:174-237.

[45] S.X.Chen,I.V.Keilegom. A review on empirical likelihood methods for regression [J]. Test.2009,18:415-447.

[46] A.B.Owen.Empirical likelihood for linear models[J].Annals of Statistics.1991,19,1725-1747.

[47] J.Qin,J.Lawless.Empirical likelihood and general eatimating equations[J]. Annals of Statistics.1994,22,300-325.

[48] A.B.Owen.Empirical likelihood ratio confidence regions[J].Annals of Statistics.1990,18:90-120.

[49] J.A.Nelder,R.A.Mead.simplex method for function minimization[J].Comput.J.7,308-313.

[50] Z.bai,J.W.Silverstein.Spectral analysis of large dimensional random matrices[M]. Beijing:Science Press.2006.

[51] G.Bernett. Probability inequalities for the sum of independent random variables[J]. Journal of the American Statistical Association.1962,57(1):33-45.

[52] 张涤新, 王志诚. 独立的无界随机变量和的概率不等式 [J]. 应用数学和力学.2001,22(5):529-533.

读硕期间发表的论文目录

张军舰，李玲玲. 修正 Berk-Jones 检验统计量在零假设下的极限分布 [J]. 广西科学, 2009,16(2): 113-116.

致　谢

三年的研究生生活即将结束，回想在读研的这三年里我得到了很多老师和同学的帮助，在毕业之际我借此机会向所有帮助过我的老师和同学表示诚挚的谢意：

首先我要感谢的是我的导师张军舰副教授，张老师是一个对工作认真负责，态度严谨的好老师，我的三年研究生学业正是有了张老师的悉心教诲才能得以顺利毕业。记得从进入研究生的第一天，张老师就教导我们应该怎样学会学习，怎样学会思考，更重要的是要学会做人，因此我学业上所取得的每一点进步都离不开张老师的悉心教诲，生活中遇到的每一个挫折都离不开张老师的耐心开导，张老师用他的宽容与耐心以及对生活的热情时刻激励我们做任何事要有信心，相信自己。能有幸成为张老师的第一届学生之一，我觉得我自己是幸运的。

其次我要感谢杨善朝教授、秦永松教授、邓国和教授、韦健老师、黎玉芳老师、梁鑫老师、唐织辉老师、罗荔龄老师等对我的教育和培养！感谢我的同门李海锋在平时的学习和生活中给我的关心和帮助，感谢所有的 07 级的同学还有我的师弟师姐们这三年来给予我的关心和帮助，和你们在一起是我感觉最开心的时刻，同时在你们身上也让我学到了很多东西，懂得了很多做人的道理，为我今后走上社会提供了帮助。愿我们大家的友情能天长地久！

最让我感谢的是我的父母、哥哥、姐姐还有我的男友，在我读研期间，如果没有他们的支持，我的学业不可能延续到今天，没有他们在我生活中给予的鼓励与关怀，我就不可能开开心心的走完三年的研究生生活，正是因为有了他们，我才可以勇敢的面对生活的挑战，勇于向更高的层次迈进！

最后，对所有在我求学生涯中给予过我关心，帮助和支持的亲人，老师，同学和朋友们说一声：谢谢，并祝你们大家工作顺利，永远开心！

尊师重道 敬业乐群