Lasso思想及算法

1、只有这么几个人在做LASSO，他们都是大牛，你可以直接GOOGLE他们的主页，看他们在这块发了什么文章。

yu bin, zhu ji, zhang tong, hui zou, yuan ming, Nicolai Meinshausen, Peter Bühlmann, Martin J. Wainwright, jianqing fan, Liza Levina, Peter Bickel,Tibshirani（Lasso的提出者）

2、统计和算法不是一回事的。举个例子吧，下面这篇文章就是统计的人发的，其中讨论到如何在GLM上运用SCAD -- LASSO衍生出来的一种惩罚函数项 -- 他们就做的很理论，他们很关心这个“算法”的理论性质，比如估计量是否趋近正太，如果是，lambda该以何种rate收敛。他们也讨论算法，但是他们对算法的要求很简单，能算出来就行。http://www.stat.umn.edu/~hzou/Papers/onestep.pdf

3、而如下这篇，讨论的基本属于同一个问题，如果将LASSO运用在GLM中的logistic回归上。但是，他们关心的是，如何算得又快又好。你看，这个时候，统计学界所陌生的nestiov都出来了。这个算法非常快，1000多个变量，也就几十秒的事情。

http://www.public.asu.edu/~jye02/Publications/Papers/fp817-Liu-KDD09.pdf

4、variable selection 是一个热点专题啊，特别是在high dimensional data analysis，

有很多新的方法( lasso, group lasso, elastic net, SCAD...)

5、R的包lars 提供了LASSO

6、请教 lasso regression 和bridge logistic regression

你可以去看一下网址“http://www-stat.stanford.edu/~tibs/lasso.html”上下载文章“Penalized regressions: the bridge vs the lasso”看一下就知道了，如果想了了解这方面更详细的信息，可加qq：381823441，他的硕士论文做的就是这方面的内容。

7、LASSO有很多令人期待的问题没有解决，所以还是有很多坑可以去填的。要想好好学习这块的话，先读那几篇state-of-the-art的文章,如下：

最基本那篇

http://www-stat.stanford.edu/~tibs/lasso.html

yuan ming 然后提出的 group lasso

http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.79.2062

zou hui的elastic net，可以同时自动选取相关程度比较大的变量（LASSO不能）

http://www-stat.stanford.edu/~hastie/TALKS/enet\_talk.pdf

zou hui 提出的adative lasso,试图避免lasso对参数的shrunkage

http://www.stat.umn.edu/~hzou/Papers/adaLasso.pdf

jianqing fan 2001 年提出的另外一种 nonconvex的惩罚项 -- SCAD

http://www.orfe.princeton.edu/~jqfan/papers/01/penlike.pdf

其学生08年时接着解决了SCAD的算法问题 -- 实际上就是一种adaptive lasso

http://www.stat.umn.edu/~hzou/Papers/onestep.pdf

这方面文献太多了，我给的这几篇是影响比较大的，你可以看看08年这篇的reference，在里面可以找到一些有意思的线索。另外，LASSO还被广泛应用在graphical model上，有兴趣可以看看这篇

http://www.stats.ox.ac.uk/~meinshau/consistent.pdf

http://arxiv.org/abs/0811.4463

http://www-stat.stanford.edu/~tibs/ftp/graph.pdf

这三篇是目前文献里公认的里程碑式的文章--关于gaussian graphical model的。

研究内容：

将Lasso应用于回归，可以在参数估计的同时实现变量的选择，较好的解决回归分析中的多重共线性问题，并且能够很好的解释结果。本项目注重对实际案例中的共线性问题利用Lasso的方法剔除变量，改进模型，并将其结果与以往变量选择的方法比较，提出Lasso方法的优势。

将Lasso应用于时间序列。将Lasso思想应用于AR(p)、ARMA(p)等模型，利用Lasso方法对AR(p)、ARMA(p)等模型中的变量选择，并给出具体的算法，随后进行模拟计算，阐明AR(p) 、ARMA(p)等模型的Lasso方法定阶的可行性。

尝试将Lasso方法应用到高维图形的判别与选择以及应用于线性模型的变量选择中，以提高模型选择的准确性。

研究意义：

随着科技的进步，收集数据的技术也有了很大的发展。因此如何有效地从数据中挖掘出有用的信息也越来越受到人们的关注。统计建模无疑是目前处理这一问题的最有效的手段之一。在模型建立之初，为了尽量减小因缺少重要自变量而出现的模型偏差，人们通常会选择尽可能多的自变量。但实际建模过程中通常需要寻找对响应变量最具有解释性的自变量子集—即模型选择(或称变量选择、特征选择)，以提高模型的解释性和预测精度。所以模型选择在统计建模过程中是极其重要的问题。

Lasso(Least absolute shrinkage and selection operator, Tibshirani(1996))方法是一种压缩估计。它通过构造一个罚函数得到一个较为精炼的模型，使得它压缩一些系数，同时设定一些系数为零。因此保留了子集收缩的优点，是一种处理具有复共线性数据的有偏估计。

　　Lasso 的基本思想是在回归系数的绝对值之和小于一个常数的约束条件下，使残差平方和最小化，从而能够产生某些严格等于0 的回归系数，得到可以解释的模型。R的Lars 算法的软件包提供了Lasso编程，我们根据模型改进的需要，可以给出Lasso算法，并利用AIC准则和BIC准则给统计模型的变量做一个截断，进而达到降维的目的。因此，我们通过研究Lasso可以将其更好的应用到变量选择中去。

研究意义：

一般地说，多元数据分析处理的对象是刻画所研究问题的多个统计指标在多次观察中呈现的数据，样本数据具有离散且有限的特征。但是，现代的数据收集技术所收集的信息，不但包括传统统计方法所处理的数据，还包括具有函数形式的过程所产生的数据。在处理数据的时候我们就会遇到模型建立的问题，这时候我们就把一些多元数据分析模型应用到函数型数据中(比如线性模型)，那么在线性模型中变量的选择问题就很重要了。

在分析这种模型的时候，人们根据问题本身的的专业理论及有关经验，常常把各种与因变量有关的自变量引进模型，其结果是把一些对因变量影响很小的，有些甚至没有影响的自变量也选入模型中。这样一来，不但计算量大，而且估计和预测的精度也会下降。此外，在一些情况下，某些自变量的观测数据获得代价昂贵，如果这些自变量本身对因变量的影响很小或根本没有影响，但我们不加选择都引到模型中，势必造成观测数据收集和模型应用费用不必要的加大。

因此，本项目基于数据的普遍特征，在对数据分析时，必须对进入模型的自变量作精心的选择。而Lasso以缩小变量集（降阶）为思想，是一种收缩估计方法。Lasso方法可以将变量的系数进行压缩并使某些回归系数变为0，进而达到变量选择的目的，可以广泛的应用于模型改进与选择。我们通过选择惩罚函数，借用Lasso思想和方法实现变量选择的目的。

国内外研究现状分析：

Tibshirani,R.(1996)在Frank(1993)提出的“Bridge Regression”和Bireman(1995)提出的“Nonnegative Garrote”的启发下提出了一种称之为Lasso (Least absolute shrinkage and seleetion operator)的新的变量选择方法并将其成功应用于COX模型的变量选择。该方法克服了传统方法在选择模型上的不足，因此该方法在统计领域受到了极大的重视。但是该方法缺乏有效的算法支撑。因此很多学者在这方面展开了研究。

Fu(1998)提出了“Shooting”算法，接着Osbome,M.R.等发现Lasso回归的解的路径是逐片线性的并提出相应的同伦算法；Bradley Efron等(2004)提出的最小角回归(Least Angle Regression)算法很好地解决Lasso的计算问题；Fan和Li(2001)指出Lasso估计对于绝对值较大的系数的压缩过大，可能会造成不必要的模型偏差，并且推测Lasso估计不具有“哲人”性质(“oracle properties”)，还给出了一种被简称为SCAD新的惩罚函数，并沿用GCv方法来选择惩罚参数(或者叫“Tunning parameters”)。

Lasso这种方法使一些回归系数变小，有些甚至被压缩为0，但忽略了相邻两个回归系数间的差异，使估计值波动性很大。在2005年，Tibshiralli和Sawnders又将此方法改进，又加入了约束条件，控制了回归系数的波动性。

模型选择本质上是寻求模型稀疏表达的过程，而这种过程可以通过优化一个“损失”十“惩罚”的函数问题来完成。

主要参考文献：

[1] Jun Liu. Large-Scale Sparse Logistic Regression[J]．KDD’09．

[2] Hui Zou and Runze Li. One-step sparse estimates in nonconxave penalized[J]．The Annals of Statistics,2008,11．

[3] Tibshirani,R.Regression Shrinkage and Selection Via the Lasso[J]．Journal of the Royal Statical Society.Series B.58,267-288．

[4] Wang Zhanfeng. A LASSO-Type Approach to Variable Selection and Estimation for Censored Regression Model[J]．2010,02．

[5] 邱南南.基于Lasso 改进的一般因果关系检验[J]．统计与信息论坛，2008,02．

[6] 赵婷婷.AR(p)模型的Lasso方法定阶[D]．东北师范大学硕士论文，2008,05．

[7] 宋国栋.线性不等式约束下的变量选择[D]．东北师范大学硕士论文，2007,05．

[8] 孙丽丽.工具变量回归模型中的变量选择[D]．东北师范大学硕士论文，2008,05．

[9] 刘小明.数据降维及分类中的流行学习研究[J]．浙江大学博士学位论文，2007,4．

[10] 杨威.函数型回归模型的成分选取[D]．东北师范大学硕士论文，2009,05．

技术路线：

根据已发表论文中遇到的违反估计的情况进行总结研究，提出补救方案，进而根据不同参数估计方法的优劣，针对实际问题给出相应的参数估计方法或是方法的组合。

    实验方案及可行性分析

第一年主要从事理论研究，通过阅读相关的书籍和学术论文，对Lasso的最新成果刻苦钻研，找出着手点，并尝试算法的改进，对不同的模型尝试借用Lasso实现变量的筛选。

  后半年主要从事Lasso在变量选择中的应用。

作者从研一开始就阅读Lasso的相关论文，通过阅读与Lasso相关的近期论文，有了一定的想法和思考，并尝试着惩罚项的添加和模拟。

2．本项目的创新之处及预期成果

创新之处：

试图将Lasso应用于现在未涉及的模型，通过添加惩罚项，利用AIC准则和BIC准则给统计模型的变量做一个截断，进而达到降维的目的。并将结合实例分析与其他模型中变量选择的方法进行比较，突出Lasso方法在多元统计分析问题中的优点。

预期成果：

（1）给出Lasso思想应用在部分模型的优劣，并在借用惩罚函数达到剔除部分变量的基础上，降低研究成本和提高模型成果的解释力度。并与传统的变量选择方法比较，突出Lasso在实际分析应用中的优势。

（2）针对实际问题和数据特征，给出相应的Lasso模型变量的选择，并与其他变量方法选择进行比较优势。

（3）预期成果：在核心期刊发表论文1-2篇。

3．与本项目有关的研究工作积累和已取得的研究成果

工作积累：

（1）通过学习多元统计分析，对变量的选择和相关的理论知识有了深入的研究。

（2）通过阅读与Lasso相关的论文，有了一定的想法和思考,并开始理论改进和案例分析。