学校编码：10384　　分类号_____密级_____
学号：15420141151998　　UDC_____

厦门大学

硕 士 学 位 论 文

超高维非参数可加模型的变量选择–基于向前可加回归算法

Variable Selection in Ultra-High Dimension Nonparametric Additive Model–Forward Additive Regression Method

李 子 豪

指导教师姓名：　钟　威　副教授
专 业 名 称：　统计学（数理统计）
论文提交日期：　2017　年　4　月
论文答辩时间：　2017　年　4　月
学位授予日期：　2017　年　6　月

答辩委员会主席：_____
评　　阅　　人：_____

2017 年　4　月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘要

本文提出向前可加回归方法(Forward Additive Regression)来解决超高维非参数可加模型的变量选择问题，超高维问题下的维数$p_n = O(\exp(n^\alpha))$。在超高维数据中，自变量的个数远大于样本量，这种“大p小n”的特性给传统的多元统计方法带来了新的挑战。当变量个数p远大于样本数n 时，传统的变量选择方法不再适用。为解决$p \gg n$下的变量选择问题，统计学者们提出了独立扫描方法(Independence screening)。独立扫描方法运用单个变量与因变量Y之间的边际得分来衡量变量的重要性，得到边际得分后通过人为给定的阈值将边际得分低于阈值的变量剔除。但独立扫描方法存在以下两个问题：1、当某个重要变量与因变量边际独立时，独立扫描方法倾向于忽略该重要变量。2、当不重要变量与强信号重要变量高度相关时，独立扫描方法很可能忽略弱信号的重要变量。Wang (2009) 针对独立扫描方法所具有的缺点提出了向前回归方法，该方法能够有效解决超高维线性模型下的变量选择问题。向前回归方法在进行变量选择时会利用已入选变量的信息，这降低了重要变量因为信号较弱而无法被选出的概率。在解决实际问题当中我们无法掌握足够的信息来确定模型的结构，Stone (1985) 提出的可加模型是一种适用范围广、弹性大的模型，为此我们将可加模型作为本文的研究对象。向前可加回归方法是向前回归方法的拓展，该方法同样能够克服独立扫描方法所具有的两个缺点。从相关的数值模拟中可以看出，向前可加回归方法在解决不同情形下的变量选择问题都能保持稳定的选择效果。

关键词：向前可加回归方法; 向前回归方法; 样条函数; 高维数据; 独立扫描方法; 变量选择

Abstract

In this paper, we propose a new method, Forward additive regression method, to solve the variable selection problem in ultra-high dimension non-parametric additive model with the dimension p_n following $p_n = O(\exp(n^\alpha))$. In ultra-high dimensional data, the number of independent variables is much larger than the sample size. The characteristic of the "$P > n$" has brought new challenges to the traditional multivariate statistical method. When the number of variables p is much larger than the sample size n, we can not use traditional variable selection methods to solve the variable selection problem. Statisticians proposed a method called Independence screening to deal with this issue. Independence screening method use a marginal score between dependent variable and independent variable to measure the significance of the independent variable. If the marginal score is lower than the threshold value, the corresponding variable will be deleted. But independence screening method has following problems: Firstly, when an significant variable is marginal independent of the dependent variable, the independence screening method tends to ignore the significant variable. Secondly, when certain insignificant variable is highly correlated with a significant variable with strong signal, it tends to ignore the significant variable with weak signal. Wang(2009) proposed a forward regression method to solve the variable selection problem in ultra high dimension linear model which overcomes the shortcomings of Independence Screening method. Forward regression method utilizes information of selected variables when conducting variable selection which reduces the probability of ignoring significant variables. In application, it is impossible for users to get enough information to determine the model structure. The additive model proposed by Stone(1985) is a widely-used model with high elasticity, so we study the additive model in this paper. Forward additive regression method is an extension of Forward regression method which also overcomes the disadvantages of the independence screening method. Based on the results of numerical simulation, we can see that forward additive regression method can keep stable performance in solving variable selection problem under different situations .

Key Words: Forward additive regression; Forward regression; Spline function; High-dimensional data; Independence screening ; Variable selection

目　录

Contents

第一章　绪论

1.1 研究背景

传统的变量选择方法并不能够解决$p \gg n$时的变量选择问题，最早解决此类问题的方法是由Fan (2008) 所提出的，他所提出的方法称为SIS方法(Sure Independence Screening 方法)。SIS 方法运用单个自变量与因变量之间的皮尔逊相关系数来衡量自变量的重要性，该自变量的皮尔逊相关系数越大证明该自变量越重要。随后针对不同模型下的独立扫描方法被提出，其中包括NIS方法、DC-SIS方法等一系列的方法。但独立扫描方法存在两个重要的问题：1、当某个重要变量与因变量边际独立时，独立扫描方法倾向于忽略该重要变量。2、当不重要变量与强信号重要变量高度相关时，独立扫描方法很可能忽略弱信号的重要变量。

Wang (2009) 提出的向前回归方法在一定程度上克服了独立扫描方法所存在的缺点，该方法通过自变量与因变量回归后所得到的残差平方和来衡量自变量的重要性，回归后残差平方和最小的自变量入选至已选变量集，已入选的变量不会被剔除。

向前回归方法主要用于解决超高维线性模型下的变量选择问题，但在解决实际问题过程中，我们常常无法掌握确切的信息来确定所要研究的模型是否具有线性结构的，简单认为所要研究的模型服从线性结构很可能会导致极大的误差。当模型结构从线性结构转变为非线性结构时，向前回归方法便不适用了，改进向前回归方法使得它能够解决非线性结构下的变量选择问题便是一个很自然的想法。由于模型结构的未知性以及实际信息的有限性，这要求我们应该采用适用范围较为广泛的模型作为所要研究的对象。Stone (1985) 所提出的形如$y = \mu + f_1(x_1) + f_2(x_2) + \ldots + f_J(x_J)$的可加模型是一个不错的选择。

本论文是在探讨用向前回归方法来解决非参数可加模型下的变量选择问题是否具有理论和实际上的可行性的基础上展开的，由于线性结构与非参数结构所呈现出的差异，向前回归方法的具体算法并不能完全照搬至非参数可加模型之下。为解决非参数结构下的变量选择问题，本论文提出了向前可加回归方法(Forward Additive Regression)，该方法继承了向前回归方法的基本思想并在该思想的基础上结合了b-spline样条，b-spline样条能够将非线性函数转换为线性函数从而更好的解决可加模型(Additive Models)结构下的变量选择问题。从数值模拟结果我们可以看出，与NIS、DC-SIS、SIRS、INIS方法相比，FAR方法具有略胜一筹的变量选择效果。

1.2 研究课题以及论文中的创新点（个人贡献）

本论文的主要研究课题是探讨向前回归方法能否解决非参数可加线性模型下的

变量选择问题，若该方法不能解决非参数可加模型下的变量选择问题，我们应如何对向前回归方法进行改进？为此，本论文的主要贡献为以下两点：1、确认了向前回归方法无法解决非参数可加模型下的变量选择问题；2、对向前回归方法进行了改进使之能够解决非参数可加模型下的变量选择问题，改进后的方法称为向前可加回归方法(Forward Additive Regression)。

向前回归方法最早是由Dohono以及Cohen (2006) 提出的，但他们在自己的文章中并没有建立起有关向前回归方法的数学理论框架，向前回归方法是否具有选择一致性等理论性质是尚未得到证实的。Wang (2009) 在他们的基础上证明了向前回归方法在参数线性回归模型下是具有良好的理论性质的，并通过相关的数值模拟来说明向前回归方法相比于其他方法(SIS 方法、LARS方法、ISIS方法) 在解决参数线性模型下的变量选择问题是具有一定优势的。本文的主要贡献在于将向前回归方法的适用范围做了一个推广，将Wang在论文中所提出的向前回归方法与样条函数结合起来变成论文中所提出的向前可加回归方法(Forward Additive Regression)，然后将向前可加回归方法运用至解决非参数可加模型下的变量选择问题。

1.3 本文主要内容以及主要结果

本文首先对已有的变量选择方法进行总结，在此基础上着重介绍了与本文有密切联系的向前回归方法以及样条函数，而本文的主题向前可加回归方法就是向前回归方法与样条函数的结合。随后对向前回归可加方法的算法进行介绍，并观察向前可加回归方法是否具有良好的变量选择效果，为此我们将通过数值模拟以及实例分析来观察向前可加回归方法的变量选择效果。

数值模拟结果显示向前可加回归方法在参数结构和非参数结构下都能够保持良好稳定的变量选择效果，在某些特殊情形下(如高度相关、近线性性、边际独立)也能取得良好的选择效果，实例分析结果也同样说明向前可加回归方法相比与其他的变量选择方法要具有一定的优越性。但是向前可加回归方法也存在相关的不足：1、FAR方法的运算次数多，耗时要比一般的变量筛选方法要长，每一个重要变量的挑选需要计算所有可能变量集所对应的残差平方和；2、选择效果依赖于入选准则的设定，具有一定的不确定性。

1.4 本文主要结构

本文具体内容如下：

第一章：介绍本文所研究问题的背景，研究问题背景包括研究目的以及解决问题过程中所采用的基本思想和方法。

第二章：对已有变量选择方法进行综述，其中包括传统的变量选择方法、高维数据下的变量选择方法、非参数情形下的变量选择方法。由于本文是在向前回归方