分类号_____ 密级 _____
U D C _____ 编号 10741

兰州财经大学
LANZHOU UNIVERSITY OF FINANCE AND ECONOMICS

硕 士 学 位 论 文

论文题目 高维变点模型自适应 Group Lasso
惩罚分位回归估计

研 究 生 姓 名：慕娟

指导教师姓名、职称：田茂再 教授

学 科、专 业 名 称：统计学 数理统计学

研 究 方 向：复杂数据分析

提 交 日 期：2020.6.8

独创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其 0 他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：慕娟 签字日期：2020.6.6

导师签名：[illegible] 签字日期：2020.6.6

关于论文使用授权的说明

本人完全了解学校关于保留、使用学位论文的各项规定，同意（选择“同意”/“不同意”）以下事项：

1.学校有权保留本论文的复印件和磁盘，允许论文被查阅和借阅，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文；

2.学校有权将本人的学位论文提交至清华大学“中国学术期刊（光盘版）电子杂志社”用于出版和编入 CNKI《中国知识资源总库》或其他同类数据库，传播本学位论文的全部或部分内容。

学位论文作者签名：慕娟 签字日期：2020.6.6

导师签名：[illegible] 签字日期：2020.6.6

Adaptive Group Lasso Penalty Quantile Regression Method for High-Dimensional Change-Point Model

Candidate :MU Juan

Supervisor:TIAN Maozai

摘要

在数据分析研究过程中，随着计算机技术的迅速发展，人们会经常遇到高维数据，这些数据不仅表现出异方差特征明显，并且预测变量被分组，例如在生物应用中，检测的基因或蛋白质可以按生物作用或生物基因分组医学途径。常见的统计分析方法，如方差分析、因子分析和基于集合的函数建模，也自然表现出变量分组。针对高维数据分析处理方法，目前相关研究文献很多且方法较为广泛。在许多应用中，所获得的数据集不仅具有高维数据特征还显示异方差状态，这时更适合考虑使用分段线性回归模型对每个数据段进行建模，而每个数据段由变点分隔。但是目前针对具有变点特征的数据，大部分研究都集中在低维数据状态下，对于高维数据变点模型关注较少。

近年来，关于变点模型和高维回归的相关文献，大多数情况下都是在具有零均值误差和有界方差条件下构建模型。另一方面，众所周知，模型中存在异常值可能会在最小二乘估计方法中引起较大误差。尤其是当误差分布并不是高斯分布或其分布尾部足够大时，在研究问题时并不清楚变点前后两个时刻误差是否发生变化，还会在检测变点时产生问题，这时更适合考虑分位数回归方法，分位回归方法在高维数据分析中具有其独特魅力。在多变点模型中，变点估计可能会影响估计量属性，研究变点模型的困难首先来自于两类参数的相关性：回归参数和变点参数。但是对于高维数据分位变点回归方法的研究较少，很多时候解决方法是先结合实际，再通过一次次实验来得到结果，这是相当麻烦的，并且当变点参数于每段估计参数相关时或变点前后误差发生变化时，这种方法太过繁复。因此为了方便实际问题中的应用，需要同时考虑变点模型的两类参数问题，且简便在高维问题中的应用。

为了研究高维且存在变点的分组解释变量其性质和过程，既要确定回归变量重要组，又要在这些组之间建立层次结构。在回归问题中，协变量可以自然分组，Group Lasso 惩罚是一种很有吸引力的变量选择方法，因为它尊重数据中的分组结构。利用高维变点分位回归，即研究多阶段模型发生变化时的变点问题，本文首先构建高维变点分位回归模型并使用自适应 Group Lasso 惩罚方法对模型的变点和系数参数进行了估计；其次对于参数估计量的渐近性和其 Oracle 性质进行研究，这涉及到相关变量组的选择问题，而不需要通过假设检验。对于在变点未

知的情况下，本文利用 SQ_τ 检验方法对变点进行检测和判断。最后通过蒙特卡罗模拟数值结果表明，与文献中其他方法相比，该方法在高维分位变点模型中具有较好性能。最后利用实际数据分析说明了该模型和方法的有效性和实用性。

关键词：Group lasso 变点回归 分位回归 高维数据 Oracle 性质

Abstract

With the rapid development of computer technology, people often encounter high-dimensional data in the research. These data not only show obvious heteroscedasticity characteristics, but also forecast variables are grouped. For example, in biological applications, the detected genes or proteins can be grouped into medical pathways by biological action or biological genes. Common statistical analysis methods, such as analysis of variance, factor analysis, and function modeling based on base set, also naturally show variable groupings. In view of high-dimensional data analysis and processing methods, there are many related research literature and methods are more extensive. In many applications, the data set obtained not only has the characteristics of high-dimensional data but also shows the state of heteroscedasticity. At this time, it is more suitable to consider the use of multiphase linear regression model to model each data segment, and each data segment is separated by change points. However, most of the current research on data with change-point characteristics is focused on low-dimensional data, and less attention is paid to high-dimensional data change-point models.

In recent years, most of the literatures about change-point model and high-dimensional regression are under the condition of zero mean error and bounded variance. On the other hand, it is well known that the existence of outliers in the model may cause large errors in the least

squares estimation method. Especially when the error distribution is not Gaussian or thick-tailed, and it is not clear whether the error changes at the two moments before and after the change point, which will cause problems when detecting the change point, it is more suitable to consider quantile regression which has its unique charm in high-dimensional data analysis. In the multiphase model, the change-point estimation may affect the properties of the estimator. The difficulty in studying the change-point model first comes from the correlation of two types of parameters: regression parameters and change-point parameters. However, there are few researches on the quantile change-point regression method for high-dimensional data. In many cases, the solution is to combine the practice first and then get the results through one experiment, which is quite troublesome. Moreover, when the change-point parameter is related to each estimation parameter or the error before and after the change point changes, this method is too complex. Therefore, in order to facilitate the application of practical problems, it is necessary to consider two kinds of parameter problems of the change-point model at the same time, and to simplify the application in high-dimensional problems.

In order to study the nature and process of group explanatory variables with high-dimension and change-points, we should not only determine the important groups of regression variables, but also establish a hierarchy between these groups. In regression problems, covariates can

be grouped naturally, and group lasso penalty is an attractive variable selection method because it respects the grouping structure in data. Using the high-dimensional change-point quantile regression, that is, to study the change-point problem when the multiphase model changes, this paper first constructs the high-dimensional change-point quantile regression model and uses the adaptive group Lasso penalty method estimates the parameters of the model's cahnge points and coefficients. Secondly, it studies the asymptotics of the parameter estimators and their Oracle properties, which involves the selection of the related variable groups, without passing the hypothesis test. When the change point is unknown, this paper uses the SQ test method to detect and judge the change point. Finally, Monte Carlo simulation results show that compared with other methods in the literature, this method has better performance in high-dimensional quantile model. Finally, the effectiveness and practicability of the model and method are illustrated by the analysis of the actual data.

Keywords：Group Lasso; change-point regression; Quantile Regression; high dimensional data; Oracle properties

目录

1 引　言

1.1 研究背景及意义

随着科学技术快速发展，计算效率不断提高，复杂及高维数据集的获取已经不仅仅停留在方便，而是产生了许多更加复杂的数据集，这使得数据的获取来源更加广泛且更加容易。这种复杂数据集现在普遍存在于各种不同的专业领域之内，其中包括天文学、生物学、经济以及社会科学等，若仅仅使用传统统计方法和工具去分析这些复杂数据，已经无法满足实际数据分析需求，尤其是当研究问题过程中面对大量高维数据。其中，数据集中观测数据的维数超过了数据集的大小时被称为高维数据集。高维数据分析的一个重要目标是将一组和响应变量相关的回归变量与大量冗余回归变量分离开。例如，在遗传学中，响应变量研究中出现了大量回归变量，但是这些回归变量中只有少数的回归变量会影响到响应变量。在最小二乘估计模型的假设背景条件下，高维数据处理分析时可以考虑使用l_1惩罚或者自适应l_1惩罚，Tibshirani[1] (1996) 介绍了该方法并称其为 Lasso 惩罚。在高维数据分析过程中使用 SCAD 和 Lasso 惩罚方法都具有选择和估计参数的优点。这两个方法将真实参数的零分量估计为零的概率趋于1，该属性也被称为稀疏性，非零分量具有一个最优估计率而且它们具有渐近正态性。同时，在可以自然地对回归变量进行分组的回归问题中，Group Lasso 惩罚是一种吸引研究人员的变量选择方法，因为它尊重数据中具有的分组结构。Group Lasso 惩罚是 Lasso 的自然扩展，并且以分组方式选择变量。当组数超过样本量大小时，大量文献研究了在高维数据回归分析中设置 Group Lasso 惩罚的参数选择和估计量属性。在提供了充分假设条件下，Group Lasso 惩罚方法选择模型的回归变量与基础模型选择结果比较具有更好的精度，并且该方法估计量具有一致性。但是，在通常情况下 Group Lasso 惩罚的参数选择不一致，并且倾向于选择模型中不重要的变量组。为了提高参数选择结果的精度，得到了一种自适应 Group Lasso 惩罚方法，该方法是自适应 Lasso 惩罚的推广，需要一个初始估计量。研究结果表明如果将 Group Lasso 惩罚估计量作为初始估计量，则在某些条件下自适应 Group Lasso 惩罚在变量组选择中是一致的。

在许多应用中，所获得的数据集不仅具有高维数据特征还显示异方差状态，

这时更适合考虑使用分段线性回归模型对每个数据段进行建模，而每个数据段由变点分隔。变点问题在低维建模中已经具有很多方法研究和实际应用，但是在高维建模中依然缺少相关文献与研究，在解决高维实际问题时分位回归将是一个很好的方法。众所周知，在高维回归模型建模过程中可以使用 Lasso 惩罚对模型进行建模，但自适应 Lasso 惩罚方法更具有自动选择和参数估计的优点。变点问题的难度无外乎就在于两类参数的同时估计，在多变点模型中，变点参数的估计可能会影响估计量属性。研究变点模型的困难首先是来自于回归参数与变点参数，这两个参数类型在构建模型时所具有的依赖性而导致相互影响使得参数估计量产生误差。近年来，变点模型和高维数据回归模型在文献中受到了很多关注，大多数情况是在具有零均值误差和有界方差的假设条件下构建模型。尽管很多学者对于分位变点回归模型的相关研究颇为丰富，比如分位变点回归在不同惩罚下估计方法的相关研究；不同分位数情况下变点回归的参数估计和变点参数的研究。但是对于高维数据下分位变点回归的相关研究相对较少，目前的解决方法大部分都是先结合实际情况，再通过一次次的实验来得到想要的结果，这个研究过程是相当麻烦的，尤其是对于变点参数和每段系数参数估计相关时，并当变点前后的误差发生了变化时，使用这种方法太过于繁复。因此为了方便实际问题中的应用，需要同时考虑有关变点模型中关于两类参数估计的问题，并且简便在高维问题分析之中的应用。Ciuperca[2] (2013) 也考虑了高维变点线性回归模型，但该模型方法是在更为严格的假设条件下进行参数估计，假设误差均值为零且有限方差，研究了自适应 Lasso 惩罚估计量。证明了估计量在每个估计段上都具有 Oracle 性质。但是，当构建模型时数据中包含异常值时，自适应 Lasso 惩罚估计量可能不会十分稳健，此外，观测数应当大于要估计的参数数。在当前的研究之中，当误差的经典条件没有发生时，可以将注意力集中在具有多个变点高维分位数回归模型上。本文研究的理论意义和现实意义如下：

(1) 理论意义：本文的思想主要来自于，在许多实际应用过程中，为了研究高维且存在变点的分组解释变量的性质和过程，既要确定回归变量的重要变量组，又要在这些组之间建立层次结构。具有分组变量的线性回归模型最常见的例子是多元方差分析。但是，在许多情况下，对于线性回归模型的理论研究，经典假设被强加于误差上：均值为零和有界方差，这在实际问题应用中并不能完全满

足。然后，如果不满足这些经典假设，或者模型具有重尾误差，那么非常有意义的方法就是分位数回归方法。此外，与经典的估计方法 (最小二乘，最小绝对偏差) 相比，模型方法在均值或中位数附近，分位数回归方法提供了非常复杂的全局洞察力。这种方法可以研究解释变量如何影响响应变量的分布。另一方面，众所周知，在构建模型时存在的异常值可能会在最小二乘估计量中引起较大的误差。尤其是当误差分布不是高斯分布且分布尾部足够大时，而且在变点模型中不清楚变点前后两个时刻的误差是否发生了变化和发生了什么变化，这便会在检测数据是否具有变点时产生问题，这时数据建模更适合考虑分位数回归建模。更为准确地来说，如果回归模型中的误差$(\varepsilon_i)_{1\leq i\leq n}$使得$P[\varepsilon_i<0]=\tau$，则$\tau$分位数回归通过函数$\rho_\tau(\varepsilon)=\sum_{i=1}^{n}\varepsilon_i[\tau I_{\varepsilon_i>0}-(1-\tau)I_{\varepsilon_i\leq 0}]$的最小化来估计分位回归的参数。$\tau=1/2$时所产生的中位数回归和$l_1$惩罚估计量，也称为最小绝对偏差 (LAD) 估计量。此外，当模型中存在着大量的回归变量时，需要同时在估计每一段的参数时用零系数消除无关的回归变量，而不是每次都通过假设检验去判断无关回归变量。此时在分位回归估计中 SCAD 和 Lasso 惩罚都具有选择和估计参数的优势。并且可以确定的是这两种方法在没有变点的模型中都具有 Oracle 属性。

因此，本文结合自适应 Lasso 惩罚和 Group Lasso 惩罚将其引入到了具有变点特征的高维变点分位回归中来进行参数估计。在回归模型中当$p>n$时，考虑使用 Group Lasso 惩罚去解决高维分位回归的参数估计权重问题，同时将该模型方法引入到高维变点问题中，在非正态误差下随着不同分位数回归的条件下回归情况和变点分布的情况进行估计，相对于其他惩罚方式其在具有分组特征的数据中据有良好的性质，在高维分组数据建模中具有实际意义。

在本文中，考虑了当变点已知时的情况然后又对变点未知时其变点位置识别的讨论和对于两种参数同时估计的理论。首先针对高维变点分位回归模型使用自适应 Group Lasso 惩罚方法对模型的变点和系数参数进行了估计，对于在变点未知的情况下，本文利用由 Qka 和 Qu[3] (2011) 提出的SQ_τ检验方法对变点进行检测和判断；其次对于参数估计量的渐近性和其 Oracle 性质进行研究；然后通过使用蒙特卡洛模拟分析研究，分别对于不同方法在不同误差情况下的参数估计结果进行了比较；最后本文利用实际数据进行分析，说明了该模型方法的有效性和

实用性。

(2) 现实意义：在许多的实际应用中，获得的数据集不仅具有高维数据特征还显示的是异方差特征，这时更适合考虑使用分段线性回归模型来对每个数据段进行建模，并且每个数据段由变点进行分隔。观测数据发生变化的模型被称为是变点模型，这些数据中变点存在的位置可能是已知的，也可能是未知的。在科学研究中，经常会遇到具有变点特征的数据集，这类数据经常出现在经济学、生物医学、遗传学等领域之内。变点线性回归模型是一个变点问题的子类问题，其中响应变量与预测变量相关的线性模型在不同的数据段上发生了变化。数据集的分割通常是基于阈值变量的未知变点，如时间或年龄，或者与数据集中观察到的其他影响变量的综合影响因素，且变点的确定会受到数据类型的影响，在时序数据中变点参数一般是时间，而在面板数据中有可能是所有自变量的综合影响结果，也有可能是某一自变量的影响因素。经济数据集是变点线性回归模型应用的主要领域，许多经济时间序列数据集可以在不同的政策规定和金融体制下收集到。因此本文对此类数据的分析提出相关尝试方案，具有一定的现实意义。

1.2 国内外研究现状

近年来，高维回归分析得到了国内外学者的广泛关注。其中稀疏惩罚最小二乘回归已经成为了分析高维复杂数据的一种广泛使用的方法。最小二乘回归可以通过各种惩罚来进行正则化，例如 Bridge Penalty (Frank 和 Friedman[4] 1993)，Lasso (Tibshirani[1] 1996)，SCAD (Fan 和 Li[5] 2001)，Elastic net (Hastie[6] 2005)，Adaptive Lasso (Zou[7] 2006) 等等。许多研究学者还考虑了最小二乘以外的回归方法来进行高维数据的分析。例如，Koenker 和 Bassett[8] (1978) 提出的分位数回归方法，由于该方法所具有的鲁棒性和对响应变量与协变量之间的关系所具有的独特见解的能力，从而在高维数据统计分析文献当中引起了广泛关注与应用。但是分位回归方法不能使用估计条件为均值函数的最小二乘回归方法来估计系数参数，经典的最小绝对偏差（LAD）回归可以被视为是分位数回归方法中的一种特殊情况，分位数回归的综合方法可以在 Koenker[9] (2005) 中找到。最近，许多关于分位数回归方法的相关研究大多数都集中在当参数数量超过了观测数量的高维数据场景中。例如，在 Wu 和 Li[10] (2009)，Belloni 和 Chernozhukov[11] (2011)，Wang 等[12] (2012)，Wang 和 De[13] (2012)文献中的高维稀疏模型中研究了l_1惩罚分

位数回归模型，其中维数大于样本量。这些文献证明了 Lasso 惩罚的分位数回归估计量在接近 Oracle 性质的情况下是一致的，并且给出了所构建模型包括真实模型的条件假设。分位回归方法的一大优点就是该方法相对于对异常值非常敏感的经典估计方法，例如最小二乘法或似然估计法，分位数回归方法提供了更为稳健的估计方法。此外，还放宽了对于误差分布的附加条件假设。在最小二乘模型的假设背景下，可以考虑使用l_1或自适应l_1惩罚方法，Tibshirani[1] (1996)介绍该方法称其为 Lasso 惩罚。SCAD 和 Lasso 惩罚都具有选择和估计参数的优点。对于没有变点的回归模型，具体的研究结果参考了文献 Wu 和 Liu[10] (2009) 中关于分位数回归模型的 SCAD 估计方法，以及 Xu 和 Ying[14] (2010) 关于中位数的 Lasso 型惩罚方法的回归估计。在 Wang 和 De[13] (2012) 的论文中对于高维数据的中位数回归，即当整体变量数大于观测数时，考虑了l_1惩罚最小绝对偏差法。Lasso 惩罚方法在过去的十年中一直都是积极研究的课题，其中 Group Lasso 惩罚是 Lasso 惩罚的自然延伸，该方法是以分组的方式进行变量选择。在许多统计建模问题和应用中都出现了选择变量组的需求。例如，在方差的多因素分析中，具有多个级别的一个因素可以由一组虚拟变量表示。在非参数可加回归模型中，每个分量都可以表示为一组基本函数的线性组合。在这两种情况下，重要因素或非参数成分的选择都等于变量组的选择。然而，Group Lasso 惩罚方法存在着估计效率低和选择不一致的问题。为了解决这些问题，学者们探讨了自适应 Group Lasso 惩罚方法，并且在理论上证明了该方法具有能够一致有效地识别真实模型，且得到的估计量可以和 Oracle 估计量一样的有效。Wei F 和 Huang J[15] (2010)，Zhang C 和 Xiang Y[16] (2016) 考虑了高维线性回归模型中的自适应 Group Lasso 惩罚估计量，研究了在高维数据条件下，Group Lasso 惩罚的回归参数选择和估计的特性。并且证明了在一定的条件下，当 Group Lasso 惩罚作为初始估计量时，自适应 Group Lasso 惩罚在组选择中是一致收敛的。关于分组选择方法和这些方法的几个应用，可以参考 Huang 等人[17] (2012)。经典地，对于回归模型，误差被假定为独立于均值零和有界方差。然后，通过最小二乘 (LS) 方法估算模型，最终当执行重要变量的自动检测时，受到 Lasso 类型惩罚的影响，但如果不满足模型误差的前后两个时刻的假设，则 LS 框架将崩溃。在这种情况下，替代方法是考虑使用 Lasso 类型惩罚的分位数回归。对于分位数回归模型，Ciuperca G[18] (2019)

考虑了分位数回归在p固定不变时和p随n变化时两种情况下自适应 Group Lasso 惩罚估计量的稀疏性与渐近正态性即 Oracle 性质。Kato K[19] (2011) 研究了高维稀疏分位数回归模型 Group Lasso 惩罚估计量的统计性质。

在存在多个变点的变点模型中，变点的估计会影响到参数估计量的性质。研究变点模型的困难就在于模型中两类参数即回归参数和变点参数的相关性影响。Ciuperca[20] (2014) 也考虑了高维变点线性回归模型，但它是在误差等于零和有界方差的严格假设条件下，并且研究了一种自适应 Lasso 惩罚估计量，证明了它在每一个片段的估计都具有 Oracle 性质。Zhang B[21] (2015) 估计高维线性回归模型中变点的数量和位置，并且估计了每个变点之间间隔中的稀疏系数。在目前的研究中，在不发生经典假设的误差条件的情况下，本文将注意力集中在具有多个变点的高维分位回归。Ciuperca G 和 Matus M[22] (2019) 证明了在一些合理的假设下，分位数 Lasso 惩罚提供的变点估计是一致的。Ciuperca G[23] (2013)，采用 SCAD 惩罚的分位数模型和采用 Lasso 型惩罚的中值回归模型，同时可以估计每一段的参数并消除无关变量，证明了对于这两种惩罚估计方法，变点估计不影响 Oracle 的性质。

在实践中，例如在方差分析的情况下，通常将其视为具有分组变量的回归线性模型。对于具有解释变量分组的模型，识别相关变量组而不是单个变量更具有意义。如果误差具有正态分布，则为了检测相关的变量组，将使用F统计检验。如果误差不是高斯误差，并且如果有更多的组，那么F统计检验是不合适的。Yuan 和 Lin[24](2006）提出了使用 Lasso 惩罚的分组变量自动选择方法。最近有几篇论文考虑了使用 Lasso 类型的惩罚进行组选择。对于固定参数空间和均值为零的模型误差假设，Nardi 和 Rinaldo[25] (2008) 建立了模型选择的一致性以及非零 Group Lasso 惩罚估计量的渐近正态性且研究了相同的估计量，其中对于正态误差的特殊情况下，协变量的数量较大。对于高维回归模型，Wei 和 Huang[15] (2010) 研究了自适应 Group Lasso 惩罚方法的变量选择和参数估计属性，但假设误差是高斯分布。仍然对于用自适应 Lasso 惩罚的误差平方和，Zhang 和 Xiang[16] (2015) 考虑了当$n \to \infty$时，p_n到无穷大时的收敛情况，并证明了参数估计量的一致性和渐近性。

国内针对自适应 Lasso 惩罚分位回归的研究有李子强等[26] (2014) 其针对面

板数据利用自适应 Lasso 惩罚分位回归方法进行了研究。同时近年在自适应 Group Lasso 惩罚方法研究与应用中，牟建波[27] (2017) 对分位回归的自适应 Group Lasso 惩罚变量选择进行了研究；牛银菊[28] (2014) 对于部分线性模型利用自适应 Group Lasso 惩罚对于参数估计及变量选择问题进行了研究分析；崔琨鹏，赵强[29] (2018) 将自适应 Group Lasso 惩罚方法引入到了 logistic 模型中，给出了拥有自适应 Group Lasso 惩罚项的 logistic 模型。

同时赖学方[30] (2017) 利用贝叶斯方法对高维数据进行了研究。对于贝叶斯方法对于 Lasso 惩罚的解释与研究有，李翰芳[31] (2013)其对面板数据的贝叶斯 Lasso 惩罚分位回归方法的研究和牟建波[32] (2017)对于贝叶斯 Group Lasso 惩罚分位回归的研究。对于变点问题的研究本文参考了龙振环, 张飞鹏, 周小英[33] (2017)，他们首先通过 Lasso 惩罚方法和广义贝叶斯信息准则确定变点参数的个数，然后再通过线性回归方法来估计变点的位置与相应的回归系数，且周小英[34] (2018) 探索了单变点和多变点情况下的逐段连续的线性分位回归模型，研究模型中变点是否存在、变点的个数以及系数参数的估计的相关统计推断问题。针对删失数据张立文等[35] (2018) 在分位回归模型及变点问题中提出一种新的检测方法。郭婧, 何幼桦[36] (2017) 使用贝叶斯方法对分位数自回归模型中系数估计及变点问题进行了分析。

1.3 文献评述

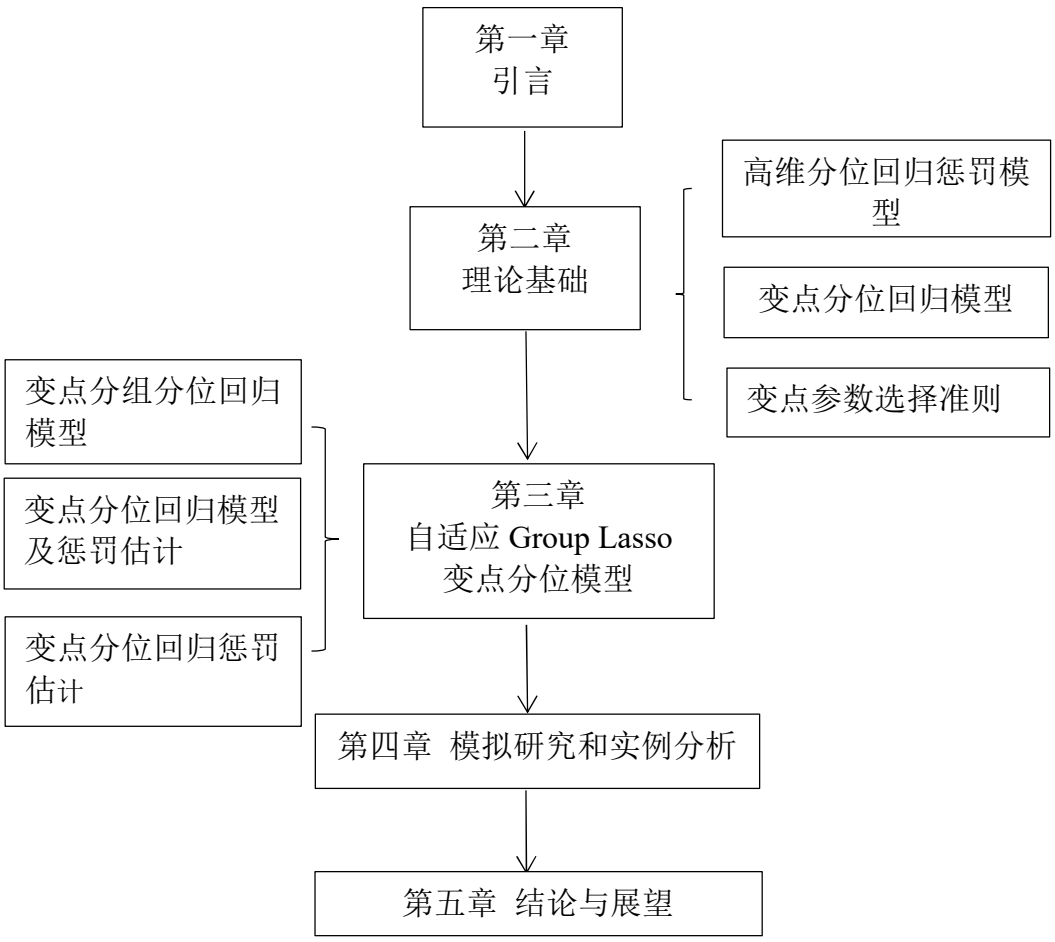
国外现有的文献大部分集中于以下几个方面：(1) 大量的文献研究具有一个变点的回归模型的性质与应用，但针对多个变点的回归模型的文献研究较少；(2) 高维变点分位回归模型的参数估计只介绍了相关性质但是针对应用方面较为欠缺；(3) 对于变点数目的评判标准水平方法比较多。

在针对国内外的相关文献的了解梳理过程中是可以发现，国内目前关于高维变点分位回归估计的文献是相对较少，并且主要集中在以下三个方面：(1) 自适应 Lasso 惩罚方法的研究，但是缺少对于自适应 Group Lasso 惩罚方法的研究；(2) 贝叶斯方法对于 Lasso 惩罚的解释与研究；(3) 变点问题有关变点的识别研究。可以发现对于高维变点分位回归的研究比较少。

1.4 文章结构

本文的文章结构内容可以分为以下的部分：第一部分主要包括目前关于高维变点分位回归模型相关的研究的发展现状以及本文所出现的创新之处，并且介绍了目前本文主题的研究背景和国内外相关研究的文献综述及文献评述，在这部分本文还提出本文主要文章结构和内容框架；第二部分是本文对基本理论知识的梳理与总结，主要是针对高维分位回归惩罚模型和变点模型进行相关学习梳理，并且简要的介绍了本文方法研究中所涉及到的一些理论基础知识，为文章后续的理论研究做铺垫；第三部分是本文研究的重点，主要是针对自适应 Group Lasso 惩罚变点分位模型进行研究分析，这部分内容主要介绍了多变点高维分位回归模型，这是本文的主要研究对象和方法创新点，还对提出的相关模型的参数估计的渐近性质的证明；第四部分是对本文所提出的方法其 Monte Carlo 模拟的研究分析与实证分析；最后部分是对于文章的总结与展望。

文章内容结构的流程框架图如下所示：



1.5 本文的创新之处

①本文将自适应 Group Lasso 惩罚方法引入到了高维变点分位回归模型之中，对两类参数进行估计并给出了估计量的大样本性质。

②本文证明了自适应 Group Lasso 惩罚方法在变点模型中的每一段参数估计的结果中都具有 Oracle 性质。

2 理论基础

本节主要是针对高维数据分析中关于高维分位回归模型的三种惩罚估计的理论知识，和变点分位回归模型及该模型在高维数据背景下两种惩罚估计方法的简要相关介绍。对于本文多次所提及到的 Oracle 性质，本文在此给出该性质的定义，便于理解该性质在本文后续证明中的意义。Oracle 估计量以及其性质是由 Fan 和 Li[5] (2001)所提出的。若已知在模型中存在无关或者弱相关的参数，对这些在研究时想要忽略的参数，使其系数为0，其中对于相关的参数，则使其系数不为零，则此时的参数选择及参数估计所得到的估计量就是 Oracle 估计量。Oracle 性质，同时在很多文献中又常常被称作是“神谕”性质，该性质一般是指下面的这两条性质：性质一，变量选择后所具有的稀疏性，该稀疏性是指对于具有真值为零的参数，其相应的估计值也为零；性质二，非零系数的渐近正态性，即对于真值不为零的参数，其相应的估计量随着样本量$n \to \infty$时，服从渐近正态性。

2.1 高维分位数回归惩罚估计

经典的高维稀疏回归模型有，

$$y_i = \beta_0 + x_{i1}^{\top}\beta_{i1} + \cdots + x_{ip}^{\top}\beta_{ip} + \varepsilon_i, i = 1, \ldots, n \tag{2.1}$$

其中，y_i是随机变量，$\boldsymbol{x}_i$是$p \times 1$独立的随机协变量向量，ε_i是独立的随机误差项，对于某个常数 $0 < \alpha < 1$，可以允许其协变量向量的维数非常大，可能大约为 $O(\exp(n^{\alpha}))$，但是只有在参数$\boldsymbol{\beta}$的分量中存在$s \ll p$的分量不为零的意义上，回归参数$\boldsymbol{\beta}$才具有稀疏性。此类模型大部分的研究点都是在如何识别出非零回归变量并且估计出其回归系数参数。由于对遗传学、机器学习、信号处理、气候异常检测、金融数据和其他的具有高维数据集特征相关新型行业的数据分析处理需求，这类模型引起了极大的关注和研究。

目前已经开发出了各种方法用来识别未知模型并且同时估计式 (2.1) 中高维稀疏模型的相应系数参数 (参考 Peng 和 Fan [28] 2004)，这些方法大部分集中在惩罚最小二乘回归方法的研究分析上。尽管其中一些方法的研究结果具有令人满意的 Oracle 性质 (Fan 和 Li [5]2001)，但是这些研究方法通常都要求模型对于不可观察到的随机误差进行严格的假设。因此，这些方法所构建的模型通

常情况下是不够稳健的，所以在实际问题应用中可能具有较差的适用性。这时研究需要考虑其他方法，与最小二乘法估计方法相比，分位数回归方法更具有鲁棒性，并且可以放宽随机误差的假设条件。分位数回归方法的优点在于：它可以提供更加完整的预测变量和响应变量之间相关关系模型；它具有出色的计算性能；并且它具有广泛的应用。将一般分位数回归模型引入到针对高维稀疏模型的l_1惩罚框架中，对于解决实际应用中发高维问题具有相当重要的意义。

2.1.1 Lasso 惩罚分位回归估计

普通线性回归模型中定义条件分位回归函数如下：

$$Q_Y(\tau|\boldsymbol{X}=\boldsymbol{x})=\boldsymbol{x}^\top\boldsymbol{\beta}_\tau \tag{2.2}$$

给定响应变量$Y\in\mathbb{R}$和协变量$\boldsymbol{X}\in\mathbb{R}^p$的向量，其中$Q_Y(\cdot)=F_Y^{-1}(\cdot)$，是给定$\boldsymbol{X}=\boldsymbol{x}$条件下响应变量$Y$累积分布函数的反函数。对于分位数$\tau\in(0,1)$存在$P(\varepsilon\leq 0|\boldsymbol{x})=\tau$，且$Q_Y(\tau|\boldsymbol{x})=\inf\{y:F_Y(y|\boldsymbol{x})\geq\tau\}$为给定分位数$\tau$的条件分位回归函数。对于给定的观测值$(\boldsymbol{x}_i,y_i)_{i=1}^n$，通过最小化$\boldsymbol{\beta}\in\mathbb{R}^p$上的经验损失函数$\sum_{i=1}^n\rho_\tau(y_i-\boldsymbol{x}_i^\top\boldsymbol{\beta})$可以获得$\boldsymbol{\beta}(\tau)$的分位数回归估计量，其中$\rho_\tau(u)=u\{\tau-I(u<0)\}$是检验函数。关于分位数回归方法参数估计量的渐近性质已得到很好的研究 (Koenker 和 Bassett[8] (1978)等)。当允许维数增加，维数$p=o(n)$时，可以使用 He 和 Shao[39] (2000) 等人的结果直接研究分位数回归估计的渐近行为。对于高维数据的情况，特别是当$p>n$时，提出了稀疏的惩罚分位数回归方法来支持系数参数估计中的稀疏性，Gu 和 Fan 等人[40] (2017)考虑将其目标惩罚函数最小化,

$$\min_{\boldsymbol{\beta}}\frac{1}{n}\sum_{i=1}^n\rho_\tau(y_i-\boldsymbol{x}_i^\top\boldsymbol{\beta})+\sum_{j=1}^n p_\lambda(|\beta_j|) \tag{2.3}$$

在此$\boldsymbol{\beta}\in\mathbb{R}^p$，$p_\lambda(\cdot),\lambda>0$是为控制模型的复杂程度而引入的惩罚函数。$p_\lambda(\cdot)$的一个常用选择是 Lasso 惩罚。Belloni 和 Chernozhukov[11] (2011)提出了l_1惩罚分位数回归及其估计量$\hat{\boldsymbol{\beta}}$：

$$Q_\tau(\beta)=\sum_{i=1}^n\rho_\tau(y_i-\boldsymbol{x}_i^\top\boldsymbol{\beta})+\frac{\lambda_n\sqrt{\tau(1-\tau)}}{n}\sum_{i=1}^p\hat{\sigma}_j|\beta_j| \tag{2.4}$$

其中$\hat{\sigma}_j = \sum_{i=1}^{n} x_{ij}^2/n, j = 1, \ldots, p$，且存在$P(\max_{1\le j\le p} |\hat{\sigma}_j - 1| \le 1/2) \ge 1 - \alpha \to 1$。在 (2.4) 式中$\lambda_n$是惩罚参数。理想情况下，惩罚函数应该具有自适应性，因为该方法可以对不重要的变量进行足够的惩罚，以迫使这部分回归系数的估计量为零，但是不会对于重要的变量进行过度惩罚，这样便可以充分识别到正确的模型，从而获得估计量 Oracle 性质。然而，通过之前的文献研究可以发现，对于 (2.4) 式中的每个变量参数的惩罚都具有相同的惩罚参数阶数λ_n/n，因此估计量就没有理想情况下所具有的适应性。所以在$\boldsymbol{\beta}(\tau)$的一些稀疏性假设条件下，Belloni 和 Chernozhukov[11] (2011) 证明了 Lasso 惩罚回归估计量具有近似的 Oracle 性质，该估计量在条件$O(\sqrt{s \log p/n})$下是一致收敛的，其中$s = \|\boldsymbol{\beta}(\tau)\|_0 = \sum_{j=1}^{p} I(\beta_j(\tau) \ne 0)$。

2.1.2 自适应 Lasso 惩罚分位回归估计

由于估计量损失函数的非平滑性质，稀疏性惩罚估计分位数回归方法在估计计算过程中具有一定的挑战性。惩罚函数的不平滑性增加了计算的复杂性，更不用说当使用惩罚函数时非凸优化计算所带来的问题了。为了改善高维稀疏模型中的分位数回归方法所遇到的问题，很多学者都尝试为不同的变量分配完全自适应的权重，并且提出了自适应l_1惩罚分位数回归估计量$\hat{\boldsymbol{\beta}}$。根据已知的文献，可以得到以下的加权l_1惩罚分位数回归，其目标函数的最小值为：

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}) + \lambda \|\boldsymbol{w}\boldsymbol{\beta}\|_1 \tag{2.5}$$

其中$\boldsymbol{w} = (w_1, \ldots, w_p)^\top$是非负权重向量，并且$w_j \ge 0, j = 1, \ldots, p$，存在$\|\boldsymbol{w}\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |w_j \beta_j| = \sum_{j=1}^{p} w_j |\beta_j|$，$\lambda > 0$是正则化参数。正则化参数$\lambda$对于自适应$l_1$惩罚分位数估计量起着至关重要的作用，它控制着自适应权重的整体大小，并对变量来进行选择，以便使无关紧要变量的回归系数估计值缩小为零，而重要变量不会受到过度惩罚。

由式 (2.3) 和 (2.5) 可以发现，自适应 Lasso 惩罚和 Lasso 惩罚之间的主要区别在于，自适应 Lasso 惩罚允许为不同变量使用不同权重的调谐参数。这样的灵敏性又对不同变量产生不同程度的收缩。换而言之，如果将相对较大的收缩量

应用于零系数，并且将相对较小的收缩量用于非零系数，则可以获得具有更好效率的估计量。在式 (2.5) 中，如果$x_{i1}=1$，且令β_1表示截距项，则可以将权重设置成$w_1=0$，这表示截距项不受到惩罚。这里需要注意的是对于 Lasso 惩罚分位数回归方法，可以选择$\boldsymbol{w}=\boldsymbol{I}_p$ (p维单位向量)。对于自适应 Lasso 惩罚分位数回归方法，通常权重采用$w_j=(|\hat{\beta}_j^{lasso}|+1/n)^{-1}, j=1,\ldots,p$，其中$\hat{\beta}^{lasso}=(\hat{\beta}_j^{lasso}, j=1,\ldots,p)^{\top}$表示分位数回归 Lasso 惩罚估计量。一旦有效地解决了 Lasso 惩罚公式即式 (2.3) 中的惩罚选择问题，便可以通过结合局部线性近似算法来高效的解决非凸惩罚分位数回归问题。局部线性近似算法解决了折叠凸惩罚分位数回归问题(Gu 和 Fan 等人[40] (2017))，

通过以下迭代：

(1) 用$\hat{\boldsymbol{\beta}}^0$表示$\boldsymbol{\beta}$的初始估计量

(2) 对于$k=1,2,\ldots,M$, 首先计算权重$w_j=\hat{w}_j^{k-1}=\lambda^{-1}p_{\lambda}^{-1}(|\hat{\beta}_j^{k-1}|)$，$j=1,\ldots,p$;然后对于式 (2.5) 中的权重问题使用计算步骤 (2) 中的权重来获得更新$\hat{\boldsymbol{\beta}}^k$。

通过上述迭代算法可以发现，惩罚分位数回归是用加权l_1惩罚分位数回归序列求解估计的。事实上，Fan 等人[41] (2014)表示在理论研究中两次或者三次的迭代结果就足够好，这样就可以得到一个高统计精度的解。例如，SCAD 惩罚具有导数，

$$p_{\lambda}^{'}(u)=\lambda I(|u|\leq\lambda)+\frac{\max(a\lambda-|u|,0)}{a-1}I(|u|>\lambda) \tag{2.6}$$

对于一些$a>2$的情况时，典型的文献选择是$a=3.7$并且只需要取$\hat{\boldsymbol{\beta}}^0=\hat{\boldsymbol{\beta}}^{lasso}$运行局部线性逼近算法进行两次迭代即可获得分位数回归的 SCAD 估计量。对于高维数据分析，自适应 Lasso 惩罚分位回归是一个良好的方法。

2.1.3 自适应 Group Lasso 分位回归估计

很多研究学者针对实际问题处理分析时，目标可能希望通过保持组来选择变量的灵活性。例如，在基因集选择问题中，生物途径可能与某个生物过程有关，但这并不一定意味着该途径中的所有基因都与该生物过程有关。这时可能不仅要

有效消除不重要的途径，而且还要鉴定重要途径中的重要基因。在此类情况下他们考虑具有p组解释变量的线性回归模型：

$$Y_i \triangleq \sum_{j=1}^{p} \boldsymbol{X}_{ij}^{\top}\boldsymbol{\beta}_j + \varepsilon_i = \mathbb{X}_i^{\top}\boldsymbol{\beta} + \varepsilon_i, i = 1, \ldots, n. \tag{2.7}$$

其中，Y_i是响应变量，ε_i是随机变量，列向量$\mathbb{X}_i$是解释变量的第i个观测值，并且包含了p组变量。对于每组j，$j = 1, \ldots, p$，$\boldsymbol{X}_{ij}$是第j组$d_j \times 1$的解释变量向量，$\boldsymbol{\beta}_j \equiv (\beta_{j1}, \ldots, \beta_{jd_j}) \in \mathbb{R}^{d_j}$是相应的第$j$组$d_j \times 1$回归系数向量。$\boldsymbol{\beta} \equiv (\beta_1, \ldots, \beta_p)$中包含所有系数的向量，对于第$i$个观测值，$\mathbb{X}_i = (\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{ip})$是所有解释变量的向量。令$\boldsymbol{\beta}_j^0 = (\beta_{j1}^0, \cdots, \beta_{jd_j}^0)$表示参数$\boldsymbol{\beta}_{j\cdot}$的未知真值，有$X_{ij,k}$表示第$j$组数据的第$k$个变量，第$i$个样本。样本可以表示为$(Y_i, \mathbb{X}_i), i = 1, \ldots, n$，相关的解释变量组对应于非零向量。当不存在惩罚时为了不失一般性，假设解释变量的第一个组$p_0(p_0 \leq p)$是相关的，并且假设当$j \leq p_0$时$\|\boldsymbol{\beta}_j^0\| \neq 0$和$j > p_0$时$\|\boldsymbol{\beta}_j^0\| = 0$ ，这里$\|\cdot\|$是指欧几里得范数。令r表示所有解释变量的数目，且$r = \sum_{j=1}^{p} d_j$。所以，p_0表示非零真参数向量的个数，r_0表示这些非零真向量中的参数总数。

针对 (2.7) 式中具有分组特征的线性回归模型可以考虑使用分位数回归方法进行分析估计，其中具有分组特征的分位数回归方法与普通分位数回归估计方法相似，都是通过最小化损失函数来得到相关的估计量，但是其在分组的选择计算上较为复杂。首先由经典的分位回归得到，分位数指数$\tau \in (0, 1)$时，并且存在检查函数$\rho_\tau(\cdot) : \mathbb{R} \to \mathbb{R}_+$，有$\rho_\tau(u) = u(\tau - I_{(u<0)})$，则参数$\boldsymbol{\beta}$的分位回归参数估计如下：

$$\tilde{\boldsymbol{\beta}}_n \equiv \arg\min_{\beta \in \mathbb{R}^r} \sum_{i=1}^{n} \rho_\tau(Y_i - \mathbb{X}_i^{\top}\boldsymbol{\beta}) \tag{2.8}$$

其相合性、$n^{-1/2}$收敛速度和估计量$\tilde{\boldsymbol{\beta}}_n^{(\tau)}$的渐近正态性已在 Koenker[9] (2005)以往的文献中得到了证明。分位数回归估计方法不进行变量自动选择，为了发现零向量组，即无关变量组，需要进行假设检验。然而，当模型 (2.7) 中p很大时，每次进行假设检验其过程繁复，自适应 Lasso 惩罚具有自动选择和参数估计的优点，不需要每次通过交叉假设检验，可以同时估计参数组和无关组。

为引入自适应 Group Lasso 惩罚估计，文献中考虑了以下参数集合：

$$A = \{j; \|\boldsymbol{\beta}_j^0\| \neq 0\} = \{1, \ldots, p_0\}$$

且$A^c = \{j; \|\boldsymbol{\beta}_j^0\| = 0\} = \{p_0 + 1, \ldots, p\}$是它的互补集。集合$A$中包含了对应于具有非零真实参数的组。在实际分析过程中，A是未知的。然后，研究目标是必须找到集合A并且估计相关的参数组。显然，当$n > p$时，普通分位回归估计是不能作为初始估计值进行计算。对于这样的模型，Group Lasso 惩罚方法是一种有吸引力的变量选择方法，因为它尊重回归变量中的分组结构。 但是在一般情况下 Group Lasso 惩罚的选择不一致，并且在模型变量选择过程中倾向于选择更多的组。为了改善参数组选择结果，考虑使用自适应 Group Lasso 惩罚方法，它是对自适应 Lasso 惩罚方法的推广。Wei 和 Huang[15] (2010)用 Group Lasso 惩罚估计量作为初始估计值，因为与普通回归模型相比，它倾向于选择更多的参数组，同时可以纠正 Group Lasso 惩罚估计的过度选择趋势，可以使用估计量性质表现良好的适应性 Group Lasso 惩罚估计值。本文利用 Wei 和 Huang[15] (2010)文中的方法，第一步使用 Group Lasso 惩罚估计值代替 (2.8) 式中所给出的估计值。即$\hat{\boldsymbol{\beta}_n}^* \equiv \arg\min_{\beta \in \mathbb{R}^r} \sum_{i=1}^{n} \rho_\tau(Y_i - \sum_{j=1}^{p} \boldsymbol{X}_{ij}^\top \boldsymbol{\beta}_j) + \lambda_n \sum_{j=1}^{p} \sqrt{d_j} \|\boldsymbol{\beta}_j\|$。

定义自适应 Group Lasso 惩罚分位回归的估计量为：

$$\hat{\boldsymbol{\beta}}_n \equiv \arg\min_{\beta \in \mathbb{R}^r} \sum_{i=1}^{n} \rho_\tau(Y_i - \sum_{j=1}^{p} \boldsymbol{X}_{ij}^\top \boldsymbol{\beta}_j) + \lambda_n \sum_{j=1}^{p} \hat{\omega}_{n;j} \|\boldsymbol{\beta}_j\| \tag{2.9}$$

权重$\hat{\omega}_{n;j} \equiv \|\hat{\boldsymbol{\beta}}_n^*\|^{-\gamma}$，$\gamma > 0$，估计量为$\hat{\boldsymbol{\beta}}_n = (\hat{\boldsymbol{\beta}}_{n;1}^\top, \ldots, \hat{\boldsymbol{\beta}}_{n;p}^\top)^\top$。在没有分组变量的分位回归情况下对于$j = 1, \ldots, p$，有$d_j = 1$时得到了 Ciuperca[20] (2014)提出和研究的自适应 Lasso 分位回归估计量。

2.2 变点分位回归

假设y_i是一个实值随机变量，$\boldsymbol{x}_i$是$p \times 1$的随机向量，$Q_{y_i}(\tau|\boldsymbol{x}_i)$是在给定$\boldsymbol{x}_i$时$y_i$的条件分位回归函数，其中$i$对应于时间指数或根据其他变量而排序的指数。这里需要注意，如果i是时间指数，则$Q_{y_i}(\tau|\boldsymbol{x}_i)$解释为$y_i$的分位回归函数，条件是由$(\boldsymbol{x}_i, y_{i-1}, \boldsymbol{x}_{i-1}, y_{i-2}, \ldots)$生成的$\sigma$代数，其中让$n$表示所有包含时间样本量。Qka 和

Qu[3] (2011) 假设条件分位回归函数在参数上是线性的，并且被K段分段结构变化所影响。

$$Q_{y_i}(\tau|\boldsymbol{x}_i)=\begin{cases}\boldsymbol{x}_i^\top\boldsymbol{\beta}_1^0(\tau) & i=1,\ldots,l_1^0,\\ \boldsymbol{x}_i^\top\boldsymbol{\beta}_2^0 & i=l_1^0,\ldots,l_2^0,\\ \vdots & \vdots\\ \boldsymbol{x}_i^\top\boldsymbol{\beta}_{K+1}^0 & i=l_K^0+1,\ldots,n,\end{cases} \tag{2.10}$$

其中$\tau\in(0,1)$表示感兴趣的分位数，$\boldsymbol{\beta}_k^0(\tau)(k=1,\ldots,K+1)$是取决于不同分位数的未知回归系数参数，而$l_k^0(k=1,\ldots,K)$是未知的变点参数。可以将$\boldsymbol{\beta}_k^0(\tau)$的子集限制在$i$上并且允许变点的存在。文献中表明回归变量$\boldsymbol{x}_i$中可以包含离散变量和连续变量。

在没有变点存在的情况下，对于给定的分位数，分位回归参数依然可以通过式 (2.8) 来求解来估计。现在假设τ分位回归模型受到K个变点的影响，其中未知的变点为$(l_1^0,\ldots,l_K^0)$。然后可能具有$l^b=(l_1,\ldots,l_K)$个变点的分位回归估计可以定义 (Qka 和 Qu[3] 2011) 为下面的函数：

$$S_l(\tau,\boldsymbol{\beta}(\tau),l^b)=\sum_{k=0}^{K}\sum_{i=l_k+1}^{l_{(k+1)}}\rho_\tau(y_i-\boldsymbol{x}_i^\top\boldsymbol{\beta}_{k+1}(\tau)) \tag{2.11}$$

这里$\boldsymbol{\beta}(\tau)=(\boldsymbol{\beta}_1(\tau)^\top,\ldots,\boldsymbol{\beta}_{K+1}(\tau)^\top)^\top$，$l_0=0$和$l_{K+1}=n$。通过文献发现可以同时估计变点参数和系数参数：

$$(\hat{\boldsymbol{\beta}}(\tau),\hat{l^b})=\arg\min_{\beta(\tau),l^b\in\Lambda_c}S_l(\tau,\boldsymbol{\beta}(\tau),l^b) \tag{2.12}$$

其中$\hat{\boldsymbol{\beta}}(\tau)=(\hat{\boldsymbol{\beta}}_1(\tau)^\top,\ldots,\hat{\boldsymbol{\beta}}_{K+1}(\tau)^\top)^\top$和$\hat{l^b}=(\hat{l}_1,\ldots,\hat{l}_K)$。具体来说，对于给定的样本，可以通过最小化$S_l(\tau,\boldsymbol{\beta}(\tau),l^b)$来估算系数$\boldsymbol{\beta}(\tau)$。然后，通过检索变点来确定最小目标函数，其中不同的变点位置对应着相关估算量最终估计值。在式 (2.12) 中Λ_c表示可能存在的变点的集合。它确保每个估计的方案都是样本的正分数。例如，可以将其指定为：

$$\Lambda_c=\{(l_1,\ldots,l_K):l_k-l_{k-1}\geq cT(k=2,\ldots,K),l_1\geq cl,l_K\leq(1-c)l\} \tag{2.13}$$

其中c是一个很小的正的常数。

2.2.1 变点分位回归的 SCAD 估计

在这节中，本文研究了 SCAD 方法在变点模型中的理论性质。对于变点不

存在的模型中，Fan 和 Li[3] (2001)对 SCAD 方法进行了理论介绍，Wu 和 Liu[10] (2009) 随后对于该方法进行了扩展。在模型 (2.10) 中，Ciuperca[2](2013) 为了研究分位回归模型的参数$(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_{K+1})$和变点参数$(l_1,\ldots,l_K)$的 SCAD 估计量，有函数：

$$S(l_1,\ldots,l_K)=\sum_{k=1}^{K+1}\inf_{(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_{K+1})\in\Gamma^{K+1}}\sum_{i=l_{k-1}+1}^{l_k}[\rho_\tau(Y_i-\boldsymbol{X}_i^\top\beta_k)+\boldsymbol{p}_{\lambda;(l_{k-1};l_k)}(|\boldsymbol{\beta}_k|)\boldsymbol{I}_p] \tag{2.14}$$

在每个间隔段(l_{k-1},l_k)中，可以考虑惩罚$\boldsymbol{p}_{\lambda;(l_{k-1},l_k)}$，其中$l_0=1$，$l_{K+1}=n$。为了简化符号的表达，可以使用$\boldsymbol{p}_{\lambda;(l_{k-1},l_k)}$来表示式 (2.14) 中$(l_{k-1},l_k)$段的惩罚项，其中$\boldsymbol{p}_{\lambda;(l_{k-1}^0,l_k^0)}$表示真实变点的惩罚。对于间隔$(1,\ldots,n)$，调谐参数$\lambda_{(0,n)}$为$\lambda_n$。因此定义了 SCAD 变点估计量为：

$$(\hat{l}_1^{(\tau,\lambda)},\ldots,\hat{l}_K^{(\tau,\lambda)})\equiv\underset{(l_1,\ldots,l_K)\in\mathbb{R}^K}{\arg\min}\ S(l_1,\ldots,l_K) \tag{2.15}$$

由 (2.14) 式中关于函数S的定义，在两个连续的变点l_{k-1}和l_k之间，相应的回归参数的 SCAD 估计量为，

$$\hat{\boldsymbol{\beta}}_{(l_{k-1};l_k)}^{(\tau,\lambda)}\equiv\arg\min_{\boldsymbol{\beta}_k}\sum_{i=l_{k-1}+1}^{l_k}[\rho_\tau(Y_i-\boldsymbol{X}_i^\top\boldsymbol{\beta}_k)+\boldsymbol{p}_{\lambda;(l_{k-1};l_k)}(|\boldsymbol{\beta}_k|)\boldsymbol{I}_p] \tag{2.16}$$

然后，通过考虑变点的相应估计量$\hat{\boldsymbol{\beta}}_{(\hat{l}_{k-1}^{(\tau,\lambda)},\hat{l}_k^{(\tau,\lambda)})}^{(\tau,\lambda)}$，来获得第$k$段的 SCAD 回归参数估计量。但是 SCAD 方法仍然具有难以在数值算法方面付诸实践的缺点。

2.2.2 变点分位回归的 Lasso 估计

一个重要的理论事实是，Lasso 惩罚估计量并没有表现出良好的 Oracle 性质。虽然 SCAD 方法所获得的估计量具有良好的 Oracle 属性，但是，该方法在数值计算时比较麻烦。因此对于分位回归模型，调谐参数λ从参数$\boldsymbol{\beta}$的一个分量变为另一个分量。因此在有变点的分位回归模型中与 SCAD 方法相比，Lasso 惩罚估计中调谐参数$\boldsymbol{\lambda}=(\lambda_{n,1},\ldots,\lambda_{n,p})$是具有不同分量的随机$p$维向量。Ciuperca[2] (2013) 考虑变点分位回归问题中对式 (2.10) 中具有K个已知变点，对于这种估计方法，其变点估计量是：

$$(\hat{l}_1^L,\ldots,\hat{l}_K^L) \equiv \underset{(l_1,\ldots,l_K)\in\mathbb{R}^K}{\arg\min} \sum_{k=1}^{K+1} \inf_{(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_{K+1})} \sum_{i=l_{k-1}+1}^{l_k} [|Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}_k| + \frac{\boldsymbol{\lambda}_{n,(l_{k-1},l_k)}^\top}{l_k - l_{k-1}} |\boldsymbol{\beta}_k|] \tag{2.17}$$

每个$k = 1, \cdots, K+1$，有第k个分段的回归参数的 Lasso 惩罚估计量为$\hat{\boldsymbol{\beta}}^L_{(\hat{l}_{k-1};\hat{l}_k^L)}$，其中$\hat{l}_0^L = 1$和$\hat{l}_{K+1}^L = n$。需要值得一提的是，对于所研究的变点存在的模型，如果在对设计矩阵的某些假设条件下，用最小二乘估计变点回归模型，并受到 Lasso 惩罚，则惩罚估计的稀疏性并不是很令人满意。 此外，由于模型中包含变点，因此在具有随机变点存在的每个区间上估计更加困难。 然后，可以考虑采用自适应 Lasso 惩罚方法来弥补这一缺点。另一方面，最小二乘自适应 Lasso 惩罚方法仅在误差均值为零且有方差的假设下成立。

这里必须强调的是，在其他的文献中已经实现的仿真在偏差，精度，识别真实零点方面均证实了自适应 Lasso 惩罚分位回归估计量所具有的优越性，尤其是在不存在误差矩或误差的中位数不为零的情况下。与其他的方法相比，在多变点的模型中，如果变点仅仅是由于误差分布所引起的而其回归参数保持不变时，则自适应分位数回归方法可提供最佳的结果。

2.3 变点参数选择准则

对于高维变点分位回归模型中存在的真实变点数K_0。回忆文献可以发现很多论文中所使用的其他的信息标准来检测变点数量。其中使用假设检验来解决线性回归中变点数量问题，用第$K+1$个变点来检测是否具有K个变点。在许多相关的研究论文中，经验似然检验被认为是可以检测到线性回归中是否具有变点存在。回顾各种文献后，对于模型变点数量的选择准则本文基于由 Oka 和 Qu[3] (2011) 所提出的SQ_τ检验。首先对该检验进行回顾，SQ_τ检验设计是在给定一个分位数指数τ的情况下去预测模型结构发生的变化：

$$SQ_\tau = \sup_{\lambda\in[0,1]} \|(\tau(1-\tau))^{-1/2}[H_{\lambda,n}(\hat{\boldsymbol{\beta}}(\tau)) - \lambda H_{1,n}(\hat{\boldsymbol{\beta}}(\tau))]\|_\infty \tag{2.18}$$

其中：

$$H_{\lambda,n}(\hat{\beta}(\tau)) = \left(\sum_{i=1}^n X_i X_i^\top\right)^{-1/2} \sum_{i=1}^{[\lambda n]} x_i \psi_\tau(Y_i - \boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}(\tau)) \tag{2.19}$$

$$\psi_\tau(u) = \tau - 1_{(u<0)}$$

对于所有的观测值中存在的变点假设$l_k^0 = [\lambda_k^0 K]$，且$0 < \lambda_1^0 < \ldots < \lambda_K^0 < 1$。关于该检测方法的非渐近干扰参数，在 Oka 和 Qu[3] (2011) 中提供了临界值表。它们不需要估计方差参数，更具体地说是模型的稀疏程度，因此即使存在多个变点，也依然存在单调性。该方法还提供了一个简单的仿真研究。其结果表明，与 Wald-based 测试相比这个检验具有更好的效果。在检测过程中仍需考虑在第K个变点的存在是否对第$K+1$个变点存在具有影响，可以表示为$SQ_\tau(K+1|K)$。将SQ_τ用于l_k到l_{k+1}段的检验，该检测通过测试每个间段是否存在额外的变点，根据公式 (2.18) 和 (2.19) 可得：

$$SQ_{\tau,k} = \sup_{\lambda\in[0,1]} \|(\tau(1-\tau))^{-1/2}[H_{\lambda,\hat{l}_{k-1};\hat{l}_k}(\hat{\boldsymbol{\beta}}_k(\tau)) - \lambda H_{1,\hat{l}_{k-1};\hat{l}_k}(\hat{\boldsymbol{\beta}}_k(\tau))]\|_\infty \tag{2.20}$$

$$H_{\lambda,\hat{l}_{k-1};\hat{l}_k}(\hat{\boldsymbol{\beta}}_k(\tau)) = \left(\sum_{i=l_{k-1}+1}^{l_k} X_i X_i^\top\right)^{-1/2} \sum_{i=l_{k-1}+1}^{[\lambda(l_k-l_{k-1})]} X_i \psi_\tau(Y_i - X_i^\top \hat{\boldsymbol{\beta}}_k(\tau)) \tag{2.21}$$

其中$\|\cdot\|_\infty$范数，例如，对于一个总体向量$z = (z_1, \ldots, z_k)$，$\|z\|_\infty = \max(z_1, \ldots, z_k)$

并且，$SQ_\tau(K+1|K)$等于$K+1$段上$SQ_{\tau,k}$的最大值，

$$SQ_\tau(K+1|K) = \max_{1\le k\le K+1} SQ_{\tau,k} \tag{2.22}$$

如果检测结果值很大，就拒绝有$K+1$个变点的模型。需要一些额外的表示来表达$SQ_\tau(K+1|K)$的极限分布。令$B_p(s)$表示$[0,1]$上p维独立布朗桥过程的向量，同时令$B_p(u,\nu) = (B_{(1)}(u,\nu), \ldots, B_{(p)}(u,\nu))'$表示独立高斯过程的$p$维向量，每个分量定义在$[0,1]^2$上且具有零均值和协方差函数。

$$E(B_{(i)}(r,u)B_{(i)}(s,\nu)) = (r\wedge s - rs)(u\wedge\nu - u\nu) \tag{2.23}$$

过程$B_{(i)}(r,u)$被称为 Brownian Pillow 或 tucked Brownian Sheet。

接下来关于确定变点个数的计算过程，不但考虑变点参数间的间隔，同时关注分位数τ。

步骤 1：对于给定的分位数指数τ，应用SQ_τ检验。如果检验没有拒绝，则得出没有变点的结论，并终止该检验过程。如果拒绝，则估计允许一个变点存在的模型。保存估计的变点，然后继续执行步骤 2。

步骤 2：开始应用$SQ_{\tau}(K+1|K)$检验从$K=1$开始。如果检验拒绝原假设，则增加K的值。在每个阶段，该模型都会被重新估计，并且变点是目标函数可以允许的K个变点的全局极小化。继续该过程，直到接受原假设。

步骤 3：$\hat{K}$表示接受原假设的的第一个值。估计有$\hat{K}$个变点的模型，保存估计的变点参数和间隔数。

3 自适应 Group Lasso 惩罚变点分位模型

在本节中,本文首先介绍具有分组变量特征的分位数回归模型。之后，给出了本文全文中使用的一些符号，然后介绍了自适应 Group Lasoo 惩罚分位回归估计量。最后，给出了关于模型误差，设计阵的一般假设。并且这部分重点给出针对变点分组分位回归模型的自适应 Group Lasso 惩罚估计方法的参数估计过程以及估计量的基本性质。在给定模型的假设条件下，并且给出估计量性质的证明过程。对于这种类型的方法，最重要的是验证是否在变点模型中保留了 Oracle 属性。因为该方法的稀疏性更具有探索意义，并且有可能受到变点估计的影响。这里想指出的是，由于惩罚方法不同，因此该结果的证明与 SCAD 估计量的证明有一些不同。

3.1 参数估计

3.1.1 变点分组分位回归模型及其假设

本文考虑具有K个变点分组分位回归的模型，即该模型中观测到的变点为$l_1,\cdots,l_K$即$1<l_1<l_2<\cdots<l_K<n$。最初，本文假设变点数量K是已知的， 但事实最常见的情况是K未知，因此本文利用已知的准则判断变点数量K。其回归模型具有以下形式：

$$\begin{aligned} Y_i=&\sum_{j=1}^{p}\boldsymbol{X}_{ij}^{\top}\boldsymbol{\beta}_{1,j}I_{(1\le i\le l_1)}+\sum_{j=1}^{p}\boldsymbol{X}_{ij}^{\top}\boldsymbol{\beta}_{2,j}I_{(l_1\le i\le l_2)}\\ &+\cdots+\sum_{j=1}^{p}\boldsymbol{X}_{ij}^{\top}\boldsymbol{\beta}_{K+1,j}I_{(l_K\le i\le n)}+\varepsilon_i, i=1,\ldots,n, \end{aligned} \tag{3.1}$$

模型 (3.1) 中参数为$(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_{K+1})=(\beta_{1,1},\ldots,\beta_{1,j},\ldots,\beta_{K+1,1},\ldots,\beta_{K+1,p})$，且变点参数为$(l_1,\ldots,l_K)$，其中$I_{(\cdot)}$是指示函数。它们的真值分别表示为$(\boldsymbol{\beta}_1^0,\ldots,\boldsymbol{\beta}_{K+1}^0)$和$(l_1^0,\ldots,l_K^0)$。观察$l_{k-1}+1,\ldots,l_k$在两个连续变点之间将称为$k$段，即第$k$个间隔。

在讨论参数$\hat{\boldsymbol{\beta}}$的主要结论之前，首先给出本文需要的假设。首先考虑确定性设计阵$(\mathbb{X}_i)_{1\le i\le n}$的相关假设，其中$\mathbb{X}_i=(\boldsymbol{X}_{i1},\ldots,\boldsymbol{X}_{ip})$是所有解释变量的向量：

假设 1 $n^{-1}\max\limits_{1\le i\le n}\mathbb{X}_i^{\top}\mathbb{X}_i\underset{n\to\infty}{\to}0$和$n^{-1}\sum\limits_{i=1}^{n}\mathbb{X}_i\mathbb{X}_i^{\top}\underset{n\to\infty}{\to}\boldsymbol{\Upsilon}$，且$\boldsymbol{\Upsilon}$是一个$r\times r$非负正

定矩阵。

关于误差ε_i的假设条件有：

假设 2 误差ε_i假设为独立同分布，且对于任意$e \in int(B), \boldsymbol{I}_r \in \mathbb{R}^r$有，其分布函数$F: B \to [0,1]$，密度分布$f$在零点附近的领域中是一致有界的。误差$\varepsilon_i$的$\tau$分位数为零，即$\tau = f(0)$

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \int_0^{\boldsymbol{X}_i^\top \boldsymbol{I}_r} \sqrt{n}[F(e+\upsilon/\sqrt{n}) - F(e)]d\upsilon = \frac{1}{2} f(e) \boldsymbol{I}_r^\top \boldsymbol{\Upsilon} \boldsymbol{I}_r \tag{3.2}$$

其中$\boldsymbol{I}_r$是r维单位向量。集合 B 是实集，且$0 \in B$。

假设 3 当$n \to \infty$时，对于调节参数λ_n和常数γ有：

$$\lambda_n \to \infty, \frac{\lambda_n}{\sqrt{n}} \to 0, n^{(\gamma-1)/2}\lambda_n \to \infty \tag{3.3}$$

假设 1 是 Lasso 惩罚方法的标准，假设 2 是分位数回归的经典假设 (Koenker[9] (2005)、Wu 和 Liu[10] (2009))。假设 1 要求设计矩阵具有合理的良好性能。对于调谐参数λ_n，适用于自适应 Lasso 惩罚分位回顾模型要求和假设 3 具有相同的条件，但未作分组解释变量。考虑到两个连续变点之间的距离，提出下列相关的假设：

假设 4 对于所有$k = 0, 1, \cdots, K$，都有$l_{k+1} - l_k \geq n^\alpha, \alpha > 1/2$。

3.1.2 自适应 Group Lasso 惩罚参数估计

为了研究自适应 Group Lasso 惩罚方法分位回归估计参数$(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{K+1})$和变点$(l_1, \ldots, l_K)$，本文考虑下面函数：

$$S(l_1, \ldots, l_K) \equiv \inf_{\substack{(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{K+1}) \\ \in \Gamma^{K+1}}} \sum_{k=1}^{K+1} \sum_{i=l_{k-1}+1}^{l_k} [\rho_\tau(Y_i - \sum_{j=1}^{p} \boldsymbol{X}_{ij}^\top \boldsymbol{\beta}_{k,j}) + \lambda_{(l_{k-1};l_k)} \sum_{j=1}^{p} \hat{\omega}_{(l_{j-1};l_j);j} \|\boldsymbol{\beta}_j\|] \tag{3.4}$$

其中$\lambda_{(l_{k-1};l_k)}$是每段的正调谐参数，每段的 Lasso 惩罚的权重为

$\hat{\omega}_{l_{j-1};l_j;j} \equiv \| \tilde{\boldsymbol{\beta}}_{(l_{j-1};l_j);j} \|^{-\gamma}$，其中$\gamma > 0$，$\tilde{\boldsymbol{\beta}}$可由式 (2.8) 获得。当$p$不随$n$发生变化时，式(3.4)中的调谐参数符合假设 3。因此对于自适应 Group Lasso 罚方法对变点估计函数有：

$$(\hat{l}_1^{(\tau,\lambda)},\cdots,\hat{l}_K^{(\tau,\lambda)}) \equiv \underset{(l_1,\cdots,l_K)\in\mathbb{R}^K}{\arg\min}\ S(l_1,\cdots,l_K) \tag{3.5}$$

回归参数的自适应 Group Lasso 惩罚分位数回归估计量为$\hat{\boldsymbol{\beta}}_{(l_{k-1};l_k)}^{(\tau,\lambda)}$，且$k=1,\cdots,K+1$。 因此，在实践过程中，首先选择估算变点，然后估算每个估算间隔的其他参数。在两个变点l_{k-1}和l_k之间的，自适应 Group Lasso 惩罚分位回归估计的相关参数$\boldsymbol{\beta}_k$估计为：

$$\hat{\boldsymbol{\beta}}_{(l_{k-1};l_k)}^{(\tau,\lambda)} \equiv \underset{\boldsymbol{\beta}_k}{\arg\min} \sum_{i=l_{k-1}+1}^{l_k} [\rho_\tau(Y_i-\boldsymbol{X}_i^\top\boldsymbol{\beta}_k)+\lambda_{(l_{k-1};l_k)}\sum_{j=1}^{p}\hat{\omega}_{(l_{j-1};l_j);j}\|\boldsymbol{\beta}_j\|] \tag{3.6}$$

通过考虑变点的相应估计量，得到第k段的自适应 Group Lasso 惩罚分位回归参数估计量：$\hat{\boldsymbol{\beta}}_{(l_{k-1};l_k)}^{(\tau,\lambda)}$。本文给出了估计量$\hat{\boldsymbol{\beta}}_{(l_{k-1};l_k)}^{(\tau,\lambda)}$的渐近性的定理，对于这种类型的方法，最重要的是验证是否将 Oracle 属性保存在变点模型中。稀疏性可能会由于变点的存在受到一定的影响。

3.2 估计参数的渐近性质

为了研究分位数回归的性质和估计量 (3.4)，首先介绍下列分位回归相关随机过程：

$$G_i^\tau(\boldsymbol{\beta};\boldsymbol{\beta}^0)=\rho_\tau(\varepsilon_i-\boldsymbol{X}_i^\top(\boldsymbol{\beta}-\boldsymbol{\beta}^0))-\rho_\tau(\varepsilon_i),$$

$$K_n^{(\tau)}(\boldsymbol{\beta};\boldsymbol{\beta}^0)=\sum_{i=1}^{n}G_i^{(\tau)}(\boldsymbol{\beta};\boldsymbol{\beta}^0)$$

$$D_i=(1-\tau)I_{(\varepsilon_i\le 0)}-\tau I_{(\varepsilon_i>0)},\ W_n=\sum_{i=1}^{n}D_i\boldsymbol{X}_i^\top$$

$$R_i^{(\tau)}(\boldsymbol{\beta};\boldsymbol{\beta}^0)=G_i^{(\tau)}(\boldsymbol{\beta};\boldsymbol{\beta}^0)-D_i\mathbb{X}_i^\top(\boldsymbol{\beta}-\boldsymbol{\beta}^0)$$

其中$\boldsymbol{\beta}_0$是参数未知真值，显而易见其中$E[D_i]=0$，K_n^τ和$R_i^{(\tau)}$关系如下：

$$\begin{aligned}K_n^{(\tau)}(\boldsymbol{\beta};\boldsymbol{\beta}^0)-E[K_n^{(\tau)}(\boldsymbol{\beta};\boldsymbol{\beta}^0)]=&\sum_{i=1}^{n}[R_i^{(\tau)}(\boldsymbol{\beta};\boldsymbol{\beta}^0)-E[R_i^{(\tau)}(\boldsymbol{\beta};\boldsymbol{\beta}^0)]]\\&+W_n(\boldsymbol{\beta}-\boldsymbol{\beta}^0)\end{aligned}$$

通过初等计算，可以得到$|R_i^{(\tau)}(\boldsymbol{\beta};\boldsymbol{\beta}^0)|\le|\mathbb{X}_i^\top(\boldsymbol{\beta}-\boldsymbol{\beta}^0)|I_{|\varepsilon|\le\boldsymbol{X}_i^\top(\boldsymbol{\beta}-\boldsymbol{\beta}^0)|}$的概率为1。通过上述关于$G_i^{(\tau)}$的定义和假设 1 中关于$\mathbb{X}_i$一致有界的假设条件可以得到不等式$K_n^{(\tau)}(\boldsymbol{\beta};\boldsymbol{\beta}^0)\le C_n\|\boldsymbol{\beta}-\boldsymbol{\beta}^0\|_2$，其中$C_n$表示不等式中不同的常数。由此已知对于每

个参数和每个τ，过程$K_n^{(\tau)}(\boldsymbol{\beta};\boldsymbol{\beta}^0)$都具有正期望，这与设计阵$\mathbb{X}_i$的假设条件无关。在假设 2 的条件下，可以发现对于所有的$\boldsymbol{\beta}\in\Gamma$，都存在$E[G_i^{(\tau)}(\boldsymbol{\beta};\boldsymbol{\beta}^0)]\geq 0$。且在已有文献中表明，在特殊情况下，即当$\tau=1/2$时，可以得到过程$K_n^{(\tau)}-E[K_n^{(\tau)}]$从而推导得到参数$\boldsymbol{\beta}^0$的邻域。通过证明，可以得到在一般情况下，即对于任何的$\tau\in(0,1)$：设有一个正序列c_n，存在$c_n\to 0$且$nc_n^2/\log n\to\infty$。在假设 1 到假设 3 的条件下，存在一个常数$C>0$，则当$\forall\varepsilon>0$，有：

$$P\left[\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}^0\|_2\leq c_n}\left|\frac{1}{nc_n^2}[K_n^{\tau}(\boldsymbol{\beta};\boldsymbol{\beta}^0)-E[K_n^{\tau}(\boldsymbol{\beta};\boldsymbol{\beta}^0)]]\right|\geq\varepsilon\right]\leq\exp(-\varepsilon^2nc_n^2C) \tag{3.7}$$

在 Ciuperca[23] (2013)中证明了有：

$$\limsup_{n\to\infty}\left(\sup_{\boldsymbol{\beta}-\boldsymbol{\beta}^0\|_2\leq c_n}\left|\frac{1}{nc_n^2}(K_n^{\tau}(\boldsymbol{\beta};\boldsymbol{\beta}^0)-E[K_n^{\tau}(\boldsymbol{\beta};\boldsymbol{\beta}^0)])\right|\right)\geq\varepsilon,a.s. \tag{3.8}$$

对于随机过程的介绍以及对分位数回归过程其性质式 (3.7) 和式 (3.8) 的描述有助于本文在接下来对于高维变点分位回归自适应 Group Lasso 惩罚估计量收敛性质的研究。通过式 (3.6) 中惩罚估计参数$\hat{\boldsymbol{\beta}}_n^{(\tau,\lambda)}$，可以得到通过自适应 Group Lasso 惩罚分位回归方法对于相关变量组选择的的指数集

$\hat{A}_n^*\equiv\{j\in\{1,\ldots,p\};\|\boldsymbol{\beta}_{n,j}^*\|\neq 0\}$，且$\hat{A}_n^{*c}$是它的补集。

本文接下来给出两个引理，这两个引理表明了集合A中分位回归的自适应 Group Lasso 惩罚估计是渐近正态的。若零参数向量已知，非零参数向量的估计量具有相同的渐近分布。引理 1 和引理 2 建立了分位回归的自适应 Group Lasso 惩罚估计的渐近正态性和稀疏性，这意味着在p不变的的情况下，该方法估计的参数仍然具有 Oracle 性质。对于引理 1 和引理 2 的具体证明可以参考文章 Ciuperca[18] (2016)。

引理 1：在假设 1 与假设 2 的两个假设条件下，可以得到$\sqrt{n}(\hat{\boldsymbol{\beta}}_n^*-\boldsymbol{\beta}^0)_A\xrightarrow[n\to\infty]{L}N(\mathbf{0}_{r^0},\tau(1-\tau)f^{-2}(0)\boldsymbol{\Upsilon}_A^{-1})$，其中$\boldsymbol{\Upsilon}_A$是$\boldsymbol{\Upsilon}$的子矩阵。

引理 2 在引理 1 和假设$n^{\gamma/2-1}\lambda_n\to\infty$的条件下，当$n\to\infty$时，可以有$\lim_{n\to\infty}P[\hat{A}_n^*=A]=1.$

给出了证明$\hat{\boldsymbol{\beta}}_n^{(\tau,\lambda)}$稀疏性所需的 Karush-Kuhn-Tucker(KKT) 最优性条件。对

于所有$j \in A_n$，都有以下等式，且概率为 1

$$\tau \sum_{i=1}^{n} X_{ij} - \sum_{i=1}^{n} X_{ij} I_{Y_i < \boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_n^{(\tau,\lambda)}} = \frac{\lambda_n \hat{\omega}_{n;j} \hat{\boldsymbol{\beta}}_n^{(\tau,\lambda)}}{\|\hat{\boldsymbol{\beta}}_n^{(\tau,\lambda)}\|} \tag{3.9}$$

对于所有$j \notin A_n$，都有概率为 1 的下列不等式

$$\left| \tau \sum_{i=1}^{n} X_{ij} - \sum_{i=1}^{n} X_{ij} I_{Y_i < \boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_n^{(\tau,\lambda)}} \right| = \lambda_n \hat{\omega}_{n,j} \tag{3.10}$$

并且引理 1 和引理 2 有助于证明自适应 Group Lasso 惩罚估计在变点分位数回归模型之中的每一段回归估计值都具有 Oracle 性质。然后，对于$n \to \infty$时，存在$(l_k - l_{k-1})^{-1} \max_{l_{k-1} \le i \le k_k} \mathbb{X}_i^\top \mathbb{X}_i \to 0$，因此可以假设对于每个间隔都存在着当$n \to \infty$时，矩阵$(l_k - l_{k-1})^{-1} \max_{l_{k-1} \le i \le k_k} \mathbb{X}_i \mathbb{X}_i^\top$收敛于$\boldsymbol{\Upsilon}_k$，其中$\boldsymbol{\Upsilon}_k$是非负定矩阵且奇异。可以用$\boldsymbol{\Upsilon}_k^0$来表示真正变点$l_k^0$的极限矩阵，$k = 1, \ldots, K$。用$\nu_{k;ij}^0$表示矩阵$\boldsymbol{\Upsilon}_k^0$中的第$i$行$j$列元素。

与自适应 Lasso 惩罚分位数回归方法相比，如果考虑考虑附加的惩罚项，则对于渐进的高斯误差假设，可得到了相同的方差矩阵。在实际应用中，集合A 是未知的，为了使估计量$\hat{\boldsymbol{\beta}}_n^{(\tau,\lambda)}$有意义，当$n$趋近于无穷大时，这两个集合必须与收敛到 1 的概率一致。 通过以下定理，可以证明自适应 Group Lasso 惩罚估计量满足了 Oracle 属性。

定理 1： 在假设 2 和假设 4 的条件下，有调谐参数$(\lambda_{(l_{k-1},l_k)})_{1 \le k \le K+1}$，是一个依赖于$n$收敛于零的序列，$(l_{k-1} - l_k)^{1/2} \lambda_{(l_{k-1},l_k)} \to \infty$和一个确定序列$(c_n)$，当$n \to \infty$时，有$c_n \to 0$，$nc_n^2 / \log n \to \infty$和$\lambda_n c_n^{-2} \to 0$，然后可以得到对于每个$k = 1, \ldots, K$都有$\hat{l}_k^* - l_k^0 = O_p(1)$。

证明：首先，根据假设 3 即式 (3.3) 和引理 2 中关于调谐参数λ的假设下自适应 Group Lasso 惩罚分位回归估计量的性质，由本节最开始给出的根据前面关于分位回归过程及其性质可以证明变点分位回归模型中所有自适应 Group Lasso 惩罚估计量都与其相应的真值之间的距离小于$n^{1/2}$。然后去证明对于每个$k = 1, \cdots, K$当$n \to \infty$时，都存在$P[|\hat{l}_k^* - l_k^0| > n^\alpha] \to 0$，即考虑一组变点，所有

的变点都靠近真实点，且距离小于$[n^{\alpha}]$。然后对于每个真实的变点l_k^0，其中$k \in \{1,\ldots,K\}$，利用式 (3.4) 中给出的函数S计算变点$l_1 \ldots l_K$，$l_1^0 \ldots l_{k-1}^0$，$l_k^0 - [n^{\alpha}]$，$l_k^0 + [n^{\alpha}]$，$l_{k+1}^0 \ldots l_K^0$且$\alpha \in (1/2, 1)$有，

$$\begin{aligned} S^*(l_1,\cdots,l_K) \geq S^*(&l_1,\cdots,l_K,\cdots,l_1^0,\cdots, \\ &l_{k-1}^0, l_k^0 - [n^{\alpha}], l_k^0 + [n^{\alpha}], l_{k+1}^0,\cdots,l_K^0) \end{aligned} \tag{3.11}$$

对于在l_{t-1}^0和l_t^0之间的观测值惩罚总和，其中$t \in \{1,\ldots,k-1,k+1,\ldots,K\}$，考虑到变点$k_{1,t} < \cdots < k_{J(t),t} \equiv \{l_1,\ldots,l_K\} \cap \{j; l_{k-1}^0 < j < l_k^0\}$。然后对于每个$t \in \{1,\ldots,k-1,k+1,\ldots,K\}$，有：

$$\begin{aligned} 0 \geq & \sum_{j=1}^{J(t)+1} \min_{\boldsymbol{\beta}_j \in \Gamma} \left[\sum_{j=k_{j-1},t+1}^{k_{j,t}} \rho_\tau(\varepsilon_i - \mathbb{X}_i^\top(\boldsymbol{\beta}_j - \boldsymbol{\beta}_t^0)) + \lambda_{(k_{j-1};k_{j,t})} \sum_{j=1}^{p} \hat{\omega}_{(l_{j-1};l_j);j} \|\boldsymbol{\beta}_j\| \right] \\ & - \sum_{j=1}^{J(t)+1} \left[\sum_{j=k_{j-1},t+1}^{k_{j,t}} \rho_\tau(\varepsilon_i) + \lambda_{(k_{j-1};k_{j,t})} \sum_{j=1}^{p} \hat{\omega}_{(l_{j-1};l_j);j} \|\boldsymbol{\beta}_j\| \right] \\ & \geq -2(K+1) \sup_{1 \leq l \leq j \leq n} \left| \inf_{\boldsymbol{\beta}} \sum_{i=l+1}^{j} G_{i,(l;j)}^{\tau,\lambda}(\boldsymbol{\beta}, \boldsymbol{\beta}^0) \right| \end{aligned} \tag{3.12}$$

其中$G_{i,(l;j)}^{\tau,\lambda}(\boldsymbol{\beta}, \boldsymbol{\beta}^0) \equiv G_i^{\tau}(\boldsymbol{\beta}, \boldsymbol{\beta}^0) + \frac{\lambda_{l,j}}{j-1} \hat{\omega}_{k_{j-1,t};k_{j,t}} \|\boldsymbol{\beta}_j\|$。存在随意的两个间隔$j$和$k$，对于$0 \leq j \leq k \leq n$，若$\lambda_{j,k} = o(n^{1/2})$，有$\lambda_{j,k} \parallel \hat{\omega}_{j,k} \parallel_2 = o_p(n^{(1+\gamma)/2})$，因此可得$\sup \left| \inf_{\boldsymbol{\beta}} \sum_{i=l+1}^{j} G_{i,(l;j)}^{\tau,\lambda}(\boldsymbol{\beta}, \boldsymbol{\beta}^0) \right| = min(o_p(n^{(1+\gamma)/2}), O_p(n^{\alpha}))$。可推出 (3.11) 式最终结果为：

$$-\min(O_p(n^{\alpha}), o_p(n^{(1+\gamma)/2})) \tag{3.13}$$

对于在$l_k^0 - [n^{\alpha}]$和$l_k^0 + [n^{\alpha}]$之间的观测值惩罚总和有

$$\min_{\boldsymbol{\beta}} \left[\sum_{i=l_r^0 - [n^{\alpha}]+1}^{l_r^0} G_{i,(l_r^0 - [n^{\alpha}];l_r^0)}^{(\tau,\lambda)}(\boldsymbol{\beta};\boldsymbol{\beta}^0) + \sum_{i=l_r^0+1}^{l_r^0+[n^{\alpha}]} G_{i,(l_r^0);l_r^0+[n^{\alpha}]}^{(\tau,\lambda)}(\boldsymbol{\beta};\boldsymbol{\beta}^0) \right] \tag{3.14}$$

再由式(3.5)中所示，其中当$c_n = c > 0$对于间隔$(l_k^0, l_k^0 + [n^{\alpha}])$，存在$\varepsilon > 0$，有$c^{-1}[n^{\alpha}]^{-1} \sum_{i=l_r^0+1}^{l_r^0+[n^{\alpha}]} G_{i,(l_r^0);l_r^0+[n^{\alpha}]}^{(\tau,\lambda)}(\boldsymbol{\beta};\boldsymbol{\beta}^0) > \varepsilon > 0$可以推导得：

$$\inf_{\boldsymbol{\beta}} \sum_{i=l_r^0+1}^{l_r^0+[n^\alpha]} G^{(\tau,\lambda)}_{i,(l_r^0);l_r^0+[n^\alpha]}(\boldsymbol{\beta};\boldsymbol{\beta}^0) \geq O_p(n^\alpha) > 0 \tag{3.15}$$

然后由由式(3.11)、(3.13)、(3.14)和(3.15)以及S^*的定义可以得到$S^*(\hat{l}^*,\ldots,\hat{l}^*_K) \leq S_0^*$，$S^*(\hat{l}^*,\ldots,\hat{l}^*_K) - S_0^* > O_p(n^\alpha) > 0$。得到对于每个$k=1,\ldots,K$，当$n\to\infty$时$P\{|\hat{l}^*_k - \hat{l}^0_k| > n^\alpha\} \to 0$，然后得每个$k=1,\ldots,K$，$\hat{l}^*_k - \hat{l}^0_k = O_p(1)$。

定理 1 得证。

对于这种自适应 Group Lasso 惩罚分位回归方法，最重要的是验证是否在变点模型中保留了 Oracle 属性。稀疏性是最有趣的，并且有可能受点估计的影响。并且由于惩罚方式的不同，因此该结果的证明也不同。

定理 2 在引理 1，假设 2 和假设 4 的条件下，对于$g>0$调节参数序列$(\lambda_{(l_{k-1},l_k)})_{1\leq k\leq K+1}$满足定理 1 且当$n\to\infty$时，$(l_{k-1},l_k)^{(g-1)/2}\lambda_{(l_{k-1},l_k)}\to\infty$我们有：

① 渐近正态性：$(\hat{l}^*_k - \hat{l}^*_{k-1})^{1/2}(\hat{\boldsymbol{\beta}}_{\hat{l}^*_{k-1};\hat{l}^*_k} - \boldsymbol{\beta}^0_k)_{A^0_k}$

$= (\hat{l}^*_k - \hat{l}^*_{k-1})^{1/2}(\hat{\boldsymbol{\beta}}_{\hat{l}^*_{k-1};\hat{l}^*_k} - \boldsymbol{\beta}^0_k)_{A^0_k}(1+o_p(1)) \xrightarrow[n\to\infty]{L} N(\mathbf{0}_{r^0}, \tau(1-\tau)f^{-2}(0)(\boldsymbol{\Upsilon}^0_{A^0_k})^{-1})$

矩阵$\boldsymbol{\Upsilon}^0_{A^0_k}$包含矩阵$\boldsymbol{\Upsilon}_k$的元素，其指数在集合$A^0_k$中。

② 稀疏性：如果对于调节参数$\lambda_{(l_{k-1},l_k)}$满足$(l_{k-1},l_k)^{(g-1)/2}\lambda_{(l_{k-1},l_k)}\to\infty$，然后对于每个$k=1,\ldots,K$，都有$\lim_{n\to\infty} P[\hat{A}^0_{n,k} = \hat{A}^*_{n,k} = A^0_k] = 1$。

证明：对于定理 2 第一部分描述得渐近正态性可以通过引理 1 和定理 1 直接得证。定理 2 的第二部分稀疏性，这里由引理 2 可知$\lim_{n\to\infty} P[A\subseteq \hat{A}^*_n] = 1$。对于$j\in\hat{A}^*_n$，和上文中已知的 KKT 最优条件得：

$$\lambda_{(\hat{l}^*_{k-1};\hat{l}^*_k)}\hat{\omega}_{(\hat{l}^*_{k-1};\hat{l}^*_k);j} < 2\left\|\sum_{i=\hat{l}^*_{k-1}+1}^{\hat{l}^*_k} \boldsymbol{X}_{ij}\right\| \leq 2\sum_{i=\hat{l}^*_{k-1}+1}^{\hat{l}^*_k}\|\boldsymbol{X}_{ij}\| = 2\sum_{i=\hat{l}^*_{k-1}+1}^{\hat{l}^*_k}\left(\sum_{k=1}^{d_j} X^2_{ij,k}\right)^{1/2}$$

由柯西—施瓦斯不等式得

$$\sum_{i=\hat{l}_{k-1}^*+1}^{\hat{l}_k^*} \frac{1}{\hat{l}_k^*-\hat{l}_{k-1}^*}\left(\sum_{k=1}^{d_j} X_{ij,k}^2\right)^{1/2} \leq \left(\frac{1}{\hat{l}_k^*-\hat{l}_{k-1}^*}\sum_{i=1}^{\hat{l}_k^*}\left(\sum_{k=1}^{d_j} X_{ij,k}^2\right)\right)^{1/2}$$

$$= \left(\frac{1}{\hat{l}_k^*-\hat{l}_{k-1}^*}\sum_{i=1}^{\hat{l}_k^*}\|\boldsymbol{X}_{ij}\|^2\right)^{1/2}$$

因此由假设 1，这里可以得到：

$$\frac{1}{\hat{l}_k^*-\hat{l}_{k-1}^*}\lambda_{(\hat{l}_{k-1}^*;\hat{l}_k^*)}\hat{\omega}_{(\hat{l}_{k-1}^*;\hat{l}_k^*);j} < 2\left(\frac{1}{\hat{l}_k^*-\hat{l}_{k-1}^*}\sum_{i=1}^{\hat{l}_k^*}\|\boldsymbol{X}_{ij}\|^2\right)^{1/2} \leq C_1 < \infty$$

$$\frac{\lambda_{(\hat{l}_{k-1}^*;\hat{l}_k^*)}\hat{\omega}_{(\hat{l}_{k-1}^*;\hat{l}_k^*);j}}{\hat{l}_k^*-\hat{l}_{k-1}^*} = \frac{\lambda_{(\hat{l}_{k-1}^*;\hat{l}_k^*)}}{(\hat{l}_k^*-\hat{l}_{k-1}^*)^{\gamma/2}\|\tilde{\beta}_{n;j}^{\gamma}\|}\frac{(\hat{l}_k^*-\hat{l}_{k-1}^*)^{\gamma/2}}{(\hat{l}_k^*-\hat{l}_{k-1}^*)}$$

$$\lim_{n\to\infty} P\left[\frac{\lambda_{(\hat{l}_{k-1}^*;\hat{l}_k^*)}\hat{\omega}_{(\hat{l}_{k-1}^*;\hat{l}_k^*);j}}{\hat{l}_k^*-\hat{l}_{k-1}^*} > A\right] = 1$$

$$\lim_{n\to\infty} P[j \in A^c \cap \hat{A}^*_{\hat{l}_k^*-\hat{l}_{k-1}^*}] = 0$$

定理 2 得证。

定理 1 和定理 2 证明了自适应 Group Lasso 惩罚方法在变点模型中的每一段参数估计的结果中都具有 Oracle 性质，且回归参数的自适应 Group Lasso 惩罚分位估计参数稀疏性不受变点影响。

4 模拟研究和实证分析

所有模拟分析都是在 R 语言环境下执行的，为了评估所提出的估计方法的性能，在本节中实现了蒙特卡罗模拟研究。首先，比较了所提出方法的结果与文献中已有的其他方法。为了判断该估计方法的性能，比较三种不同估计方法估计的精准度，并且在不同的误差情况下进行比较分析。然后利用房地产数据，将时间作为变点参数，利用该方法判断出不同分位数的情况下不同变点的存在位置，结合实际背景分析政策。

4.1 Monte Carlo 模拟研究

在实际 R 语言运行过程中经过数值转换后，使用 Group Lasso 惩罚方法代替自适应 Group Lasso 惩罚方法。然后，为了计算自适应 Group Lasso 惩罚最小二乘估计，使用了 R 软件包 grpreg 中的函数 grpreg，并使用 BIC 准则在网格上选择了调谐参数。为了计算自适应 Group Lasso 惩罚分位数回归估计，本文使用了 R 软件包 rqPen 中的函数 cv.rq.group.pen，并且利用 BIC 准则进行调谐参数值的选择。

在模拟实施过程中，使用了 QICD 算法来求解分组分位回归估计，并应用 BIC 准则来选择最优正则化参数。模拟次数为 200 次，考虑了三种误差模式：标准正态分布、指数分布和柯西分布，并且比较了不在同误差分布下分位回归的 Group Lasso (Qglasso) 惩罚和自适应 Group Lasso (Qaglasso) 惩罚的两种方法，还有最小二乘回归估计的自适应 Group lasso (aglasso) 惩罚的性能。为了评估估计量的表现，给出了零系数正确估计为零的比率 (%True0) 和非零系数错误估计为零的比率 (% False 0)。

模拟研究中，有 10 个组，每个组由 5 个协变量组成。协变量向量 $X=(X_1,\ldots,X_{10})$，其中 $X_j=(X_{5(j-1)+1},\ldots,X_{5(j-1)+5}), 1\leq j\leq 10$。为了生成 X，首先模拟生成 50 个随机变量，$R_1,\ldots,R_{50}$ 独立分布为 $N(0,1)$。然后，$Z_j, j=1,\cdots,10$，用多元正态分布模拟生成，其中均值为零，方差 $cov(Z_{j1},Z_{j2})=0.5^{|j_1-j_2|}$。因此协变量 $X_1,\ldots,X_{50}$ 生成为

$$X_{5(j-1)+k}=\frac{Z_j+R_{5(j-1)+k}}{\sqrt{2}}, 1\leq j\leq 10, 1\leq k\leq 5.$$

根据提出的三种随机误差，响应变量Y由$\sum_{k=1}^{10}\boldsymbol{X}_k'\boldsymbol{\beta}_k+\varepsilon$生成，假设变点个数已知：该模型包含两个变点 (三个阶段)。参数$\boldsymbol{\beta}$设计如下：

$$\boldsymbol{\beta}_1=(\boldsymbol{\beta}_{1,1},\boldsymbol{\beta}_{1,2},\ldots,\boldsymbol{\beta}_{1,10})=(2,2,4,1,-3,1,2,1,-0.5,1,0,\ldots,0)$$

$$\boldsymbol{\beta}_2=(\boldsymbol{\beta}_{2,1},\boldsymbol{\beta}_{2,2},\ldots,\boldsymbol{\beta}_{2,10})=(0.5,1,1.5,1,0.5,1,1,1,1,1,0,\ldots,0)$$

$$\boldsymbol{\beta}_3=(\boldsymbol{\beta}_{3,1},\boldsymbol{\beta}_{3,2},\ldots,\boldsymbol{\beta}_{3,10})=(1,1,4,1,-3,1,2,1,2,1,0,\ldots,0)$$

表 4.1 当$\varepsilon_i\sim N(0,1)$时三种方法模拟对比

n	p	r	模型	% True 0	% False 0
200	10	50		10	40
			agLasso	0.97	0
	$(1,l_1)$		QgLasso	0.91	0.15
			QagLasso	0.97	0.1
			agLasso	0.81	0
	(l_1,l_2)		QgLasso	0.96	0.1
			QagLasso	0.84	0.1
			agLasso	0.81	0
	(l_2,n)		QgLasso	0.82	0
			QagLasso	0.99	0

表 4.1 是在变点已知时，当误差分布为正态分布，三种方法模拟的对比结果，其中根据零系数正确估计为零的比率 (%True0) 和非零系数错误估计为零的比率 (% False 0)，可以发现最小二乘回归方法的自适应 Group lasso (aglasso) 惩罚估计的效果最好，该方法的非零稀疏错误估计的比率在三个阶段中都为零，对于相关参数的判断更为准确。

表 4.2 当$\varepsilon_i \sim E(1)$时三种方法模拟对比

n	p	r	模型	% True 0	% False 0
200	10	50		10	40
			agLasso	0.91	0
	$(1,l_1)$		QgLasso	0.81	0.1
			QagLasso	0.82	0
			agLasso	0.81	0
	(l_1,l_2)		QgLasso	0.96	0
			QagLasso	0.95	0
			agLasso	0.81	0
	(l_2,n)		QgLasso	0.91	0
			QagLasso	1	0

在表 4.2 中，对比了误差分布呈指数分布的三种方法在每个间隔中正确和错误估计零系数的百分比。可根据模拟结果发现自适应 Group lasso 惩罚分位回归(Qaglasso)估计零系数正确估计为零的比率相较于其他两种更好。

表 4.3 当$\varepsilon_i \sim C(0,2)$时三种方法模拟对比

n	p	r	模型	% True 0	% False 0
200	10	50		10	40
			agLasso	0.88	0.2
	$(1,l_1)$		QgLasso	0.99	0.2
			QagLasso	0.97	0.1
			agLasso	0.94	0
	(l_1,l_2)		QgLasso	0.97	0.2
			QagLasso	0.97	0.16
			agLasso	0.81	0.1
	(l_2,n)		QgLasso	0.91	0.2
			QagLasso	0.91	0.16

表 4.3 给出了误差分布呈柯西分布的三种方法在每个间隔中正确和错误零的百分比。根据上述三个表，可以发现在已知两个变点三个阶段的条件下，三种不同的误差分布的估计结果。当误差分布为正态分布时，最小二乘自适应惩罚估计结果最好，但当误差分布为指数分布和柯西分布时，自适应 Group Lasso 惩罚分位回归估计的零系数正确估计为零的百分比结果较好。三种方法的估计偏差和精度的比较在后续表 4.5 中给出。

总之，通过自适应 Group Lasso 惩罚分位数和 Group Lasso 惩罚方法，变点估计不会影响回归参数估计的稀疏性。另一方面，用于非零回归参数的自适应 Group Lasso 惩罚分位数回归估计量比相应的 Group Lasso 惩罚估计量更准确。对于高斯误差，自适应 Group Lasso 最小二乘估计方法可以很好地识别n个变量的两个 (相关和不相关) 变量组。但是对于指数分布和柯西分布的有偏分布自适应 Group Lasso 分位回归估计的结果更好。

表 4.4 当变点前后不同误差三种方法模拟对比

误差分布	%true 0			%false 0		
	agLasso	QgLasso	QagLasso	agLasso	QgLasso	QagLasso
$(1, l_1)$ $\varepsilon_1 \sim E(1)$	0.84	0.86	0.96	0.13	0.12	0.02
(l_1, l_2) $\varepsilon_2 \sim N(0,1)$	0.97	0.99	0.99	0.05	0.005	0
(l_2, n) $\varepsilon_3 \sim E(1)$	0.99	0.67	0.98	0.04	0	0
$(1, l_1)$ $\varepsilon_1 \sim E(1)$	0.72	0.85	0.97	0.12	0.14	0.01
(l_1, l_2) $\varepsilon_2 \sim C(0,2)$	0.22	0.88	0.81	0.07	0.05	0.03
(l_2, n) $\varepsilon_3 \sim E(1)$	0.96	0.65	0.99	0.05	0.007	0
$(1, l_1)$ $\varepsilon_1 \sim N(0,1)$	0.86	0.98	0.97	0.08	0.04	0.02
(l_1, l_2) $\varepsilon_2 \sim E(1)$	0.90	0.80	0.99	0.02	0.005	0
(l_2, n) $\varepsilon_3 \sim N(0,1)$	1	0.99	0.99	0.04	0	0
$(1, l_1)$ $\varepsilon_1 \sim E(1)$	0.97	0.84	0.97	0.11	0.10	0.01
(l_1, l_2) $\varepsilon_2 \sim E(1)$	0.92	0.77	0.99	0.02	0.003	0
(l_2, n) $\varepsilon_3 \sim N(0,1)$	1	0.99	0.99	0.2	0	0

在表 4.4 中，通过模拟，本文得到了三种方法在每个间隔中正确和错误零的百分比，并且每个阶段的误差分布发生改变。在表中，给出三种方法的变点回归模型估计值的中位数。为了简化表述，三种方法 Group Lasso 惩罚、自适应 Group Lasso 惩罚和最小二乘回归的自适应 Group lasso 惩罚依次表达为 Qglasso、Qaglasso 和 aglasso。在模拟过程中设置正态分布为$N(0,1)$，指数分布为$Exp(1)$和柯西分布$C(0,2)$。

对于误差分布呈指数分布或柯西分布的阶段，自适应 Group Lasso 惩罚分位数回归方法可根据真零零百分比，非零参数的偏差和精度提供最佳结果。请注意，在所有情况下，当回归参数从一个阶段到另一个阶段都不相同时，通过三种方法，当分位数为中值即$\tau = 0.5$时，估计值重合或非常接近真实值。根据表 4.4，我们得出结论，如果最小二乘模型的自适应 Group Lasso 方法对于具有相同回归参数和不同指数误差分布的两个阶段之间，则得出的变点估计值不太准确。通常，对于具有自适应 Group Lasso 惩罚的最小二乘模型，在具有高斯误差的估计中可获得良好的结果，而对于指数分布或柯西分布，则获得相对较差结果。通过 Group Lasso 惩罚方法，具有指数分布的误差其估计不佳。由此本文得出结论，通过自适应 Group Lasso 分位回归方法，当误差发生变化时，对比前面误差不发生变化的情况，真实零点的选择百分比会略有下降，但是相对于其他两种方法，该方法更好，尤其是对于非零回归参数的选择。

表 4.5 当变点前后不同误差三种方法估计量精度比较

误差分布	$mean\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\|$			$1/N \parallel \hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0 \parallel_2^2$		
	agLasso	QgLasso	QagLasso	agLasso	QgLasso	QagLasso
$\varepsilon_1,\varepsilon_3 \sim E(1)$ $\varepsilon_2 \sim N(0,1)$	0.64	0.38	0.16	9.8	5.7	0.93
$\varepsilon_1,\varepsilon_3 \sim E(1)$ $\varepsilon_2 \sim C(0,2)$	4.8	0.49	0.26	41	7.5	2.7
$\varepsilon_1,\varepsilon_3 \sim N(0,1)$ $\varepsilon_2 \sim E(1)$	0.64	0.28	0.16	9.5	3	0.96
$\varepsilon_1,\varepsilon_2 \sim E(1)$ $\varepsilon_3 \sim N(0,1)$	0.64	0.26	0.16	9.9	2.87	0.94

表 4.6 三种方法变点检测对比

误差分布	$\hat{l_1}$			$\hat{l_2}$		
	agLasso	QgLasso	QagLasso	agLasso	QgLasso	QagLasso
$\varepsilon_1,\varepsilon_3 \sim E(1)$ $\varepsilon_2 \sim N(0,1)$	61	60	61	120	120	120
$\varepsilon_1,\varepsilon_3 \sim E(1)$ $\varepsilon_2 \sim C(0,2)$	61	61	61	120	120	120
$\varepsilon_1,\varepsilon_3 \sim N(0,1)$ $\varepsilon_2 \sim E(1)$	60	60	60	120	120	120
$\varepsilon_1,\varepsilon_2 \sim E(1)$ $\varepsilon_3 \sim N(0,1)$	60	60	60	120	120	120

在表4.5中计算每个估计值的偏差，即估计值与真实值之差的绝对值的均值，和估计值方差$1/N \parallel \hat{\beta}-\beta 0 \parallel_2^2$的近似值，$N=100$为蒙特卡罗模拟重复次数。总之，通过自适应 Group Lasso 惩罚分位数回归方法和 Group lasso 惩罚分位回归方法比较，可以发现变点估计不会影响回归参数估计的稀疏性。另一方面，用于非

零回归参数的自适应 Group Lasso 惩罚分位数回归估计量比相应的 Group Lasso 惩罚估计量更准确（在偏差和方差方面），其为非零参数的偏差和精度提供最佳结果。表 4.6 使用在三种估计方法的条件下判断变点估计，在模拟中，本文令第一个变点参数的真实值$l_1^0 = 60$，第二个变点参数的真实值$l_2^0 = 120$。可以发现三种方法变点判断的效果均较为良好，但是在第一段和第三段间隔误差分布为指数分布，第二间隔误差分布位柯西分布时，变点判断不准确。

4.2 案例分析

4.2.1 北京房价数据

为了防止房价增长速度过快，控制房地产市场，自 2010 年北京市开始房价“限购”等一系列政策调整市场。本文使用 2009 年至 2017 年北京住房价格，该数据来自链家网。利用自适应 Group lasso 惩罚分位回归方法对该组数据进行变量选择以及建模分析。研究各变量与价格之间的关系，其中将时间作为变点参数。这组数据中包含 792 条房屋交易数据，是 2009 年至 2017 年的季度数据。房产交易数据如下：

Y：是每平方米价格的对数；

Square：房屋总平方米的对数；

TradeTime：是交易时间；

BedRoom：卧室的数量；

Bathroom：卫生间的数量；

BuildingType：建筑类型 (建塔楼(1)，平房(2)，建板塔结合(3)，建板楼(4))

ConstructionTime：施工时间；

RenovationCondition：改造条件 (其他(1)，毛坯(2)，简装(3)，精装(4))：

ladderRatio：梯子比率，它描述了同一楼层的居民数量与梯子电梯数量之间的比例，居民平均有多少梯子。

Elevator：电梯 (有(1)，没有(0))

fiveYearsProperty：是否满五年的产权 (有(1)，没有(0))，房子满两年和满五年之间最大的区别就是税费上的不同,可以享受税费优惠政策，可以帮购房者省下了不少钱。

Subway：地铁(有(1)，没有(0))

4.2.2 数据分析

首先对数据进行初步简单分析，根据时间参数，分析 2009 年至 2017 年北京房价基本变化趋势。值得注意的是，此处分析，直接使用的是北京房屋每平方米的价格而不是每平方米价格的对数。图 4.1 是北京房屋每平方米价格随时间变化的箱线图：

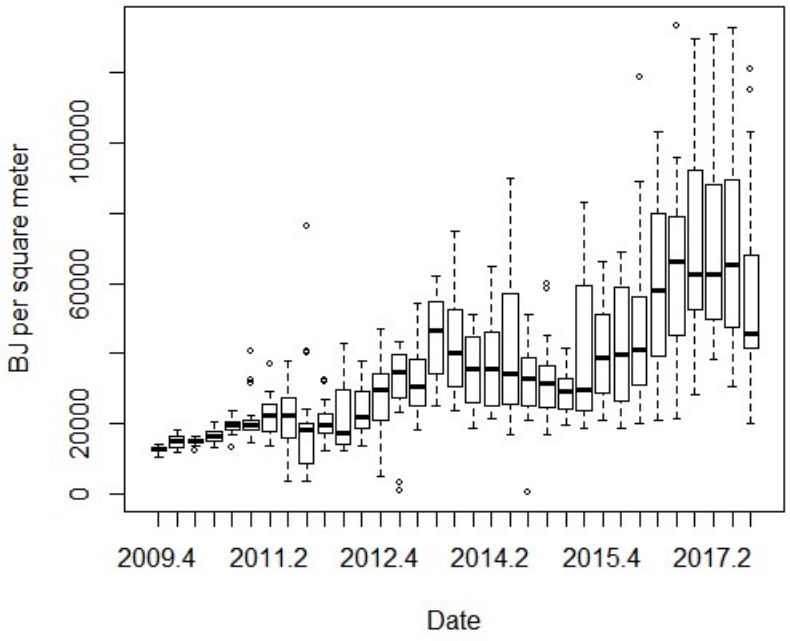


图 4.1 北京房屋每平方米价格

如图 4.1 所示，可以从 2009 年至 2017 年的数据中发现北京房价整体走向呈上升趋势，但是由于这几年北京市房产政策调控的原因导致北京房价局部时间变化明显，通过该箱线图可以明显观察到北京在 2009 年到 2017 年房价变化幅度波动，其中 2011、2013 和 2016 年价格趋势明显发生变化。

4.2.3 变点选择

本节利用SQ_{τ}检验方法对不同分位数情况下的变点进行检验。表 4.7 中是在五种不同分位数情况下检测该组数据中存在的变点情况，且判断出了这些变点存

在的位置。这里需要注意的是，为了简便运算过程，本文将房屋每平方米价格对数化。

表 4.7 不同分位数变点检测

分位数	0.20	0.35	0.50	0.65	0.80
$SQ(1\|0)$	4.856	5.968	6.733	6.601	6.519
$SQ(2\|1)$	3.137	5.038	5.239	5.454	4.461
$SQ(3\|2)$	2.527	2.717	3.235	3.002	2.140
变点数目	3	3	3	3	3
变点 1	2011.2	2011.3	2011.2	2011.3	2011.3
变点 2	2013.1	2012.4	2012.4	2012.4	2012.4
变点 3	2016.2	2016.1	2016.1	2016.1	2016.1

根据表 4.7 中所确定的变点位置，通过结合北京房价政策本文可以分析出现变点的原因和不同分位数情况下房屋价格对于不同房价政策的敏感程度，其中在不同分位点处第一个变点是 2011 年的第二和第三季度。对比该阶段政策，发现在 2010 到 2011 年北京限购限贷，并且五次加息，在 2011 年 2 月规定外地满五年限购一套，北京户籍只能拥有两套。第二个变点 2012 年第一季度和 2013 年第一季度，此时北京首次将二手房纳入调控按个人转让住房所得的 20%征收所得税，满五唯一免征个税，禁止京籍单身人士购买二套房。虽然北京房价政策调控不断，但是可以明显发现在 2016 年期间利好政策，北京房屋价与量齐涨。

5 总结与展望

5.1 结论

本文主要针对高维变点且具有分组特征的数据，研究的主要内容包括三部分，第一部分是基于分位回归模型，利用已知的惩罚方法及性质提出自适应 Group Lasso 惩罚方法，对数据进行降维并且对分位变点回归模型参数进行选择与估计，使用SQ_τ检验判断不同分为点处的变点位置；第二部分是由本文提出的惩罚方法对模型进行参数估计，并且给出了渐近性证明；第三部分是分析该方法在不同误差条件下模型的参数估计，通过模拟和实例分析可以发现，模拟结果更加真实，此方法估计效果良好。通过上述三部分理论研究，本文得出三点结论，第一点对于高维变点且具有分组特征的数据，使用自适应 Group Lasso 惩罚分位回归方法效果更加贴切，有更好的估计结果；第二点通过三种参数估计方法比较发现，在非正态误差条件下自适应 Group Lasso 惩罚估计更加精准；第三点是通过估计参数的渐近性证明，可以得出此模型的参数估计结果具有 Oracle 性质，理论上此模型的参数估计方法比较可靠，并且通过 Monte Carlo 模拟和实例分析结果更加证明了这一点。

5.2 展望

在本文的研究过程中发现在变点前后不同误差分布情况下不同的分位点的回归结果有明显差异，但是本文缺乏对于不同分位点的的结果讨论。在估计变点参数和系数参数时，在计算过程中本文使用了已有算法，没有进行新的算法创新，因此整个参数估计过程在高维数据下计算较为缓慢。此外，有兴趣的同学也可以尝试使用贝叶斯方法进行数据降维和两类参数的估计，并且在一些实例分析中，需要分析不同分位点的信息时，可以扩展分位回归方法对模型进行分析。

参考文献

[1] Tibshirani R. Regression Shrinkage and Selection Via the Lasso[J]. Journal of the Royal Statistical Society Series B (Methodological), 1996, 58(1):267-288.

[2] Ciuperca G. Quantile regression in high-dimension with breaking[J]. Journal of Statistical Theory & Applications, 2013.

[3] Oka T, Qu Z. Estimating structural changes in regression quantiles[J]. Journal of Econometrics, 2011, 162(2):248-267.

[4] Frank L E, Friedman J H. A Statistical View of Some Chemometrics Regression Tools[J]. Technometrics, 1993, 35(2):109-135.

[5] Fan J, Li R. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties[J]. Publications of the American Statistical Association, 2001, 96(456):1348-1360.

[6] Hastie Z T. Regularization and Variable Selection via the Elastic Net[J]. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 2005, 67(2):301-320.

[7] Zou, Hui. The Adaptive Lasso and Its Oracle Properties[J]. Journal of Industrial and Management Optimization (JIMO), 2006, 101(476):1418-1429.

[8] Koenker R. and Bassett G. Regression Quantiles[J]. Econometrica,1978, 46:33-50.

[9] Koenker R. Quantile Regression[J]. 2005, 15(4):143-156.

[10] Wu Y, Liu Y. Variable selection in quantile regression[J]. Stata Sinica, 2009, 19(2):801-817.

[11] Belloni A, Chernozhukov V. l_1-penalized quantile regression in high-dimensional sparse models[J]. The Annals of Statistics, 2011, 39(1):82-130.

[12] Wang L, Wu Y, Li R. Quantile Regression for Analyzing Heterogeneity in Ultra-High Dimension[J]. Journal of the American Statistical Association, 2012, 107(497):214-222.

[13] Wang L, De Leeuw J. The L_1 penalized LAD estimator for high dimensional linear regression[J]. Journal of Multivariate Analysis, 2012, 120(9):135-151.

[14] Xu J, Ying Z. Simultaneous estimation and variable selection in median regression using Lasso-type penalty[J]. Annals of the Institute of Statistical Mathematics, 2010, 62(3):487-514.

[15] Wei F, Huang J. Consistent group selection in high-dimensional linear

regression[J]. Bernoulli, 2010, 16(4):1369-1384.

[16] Zhang C, Xiang Y. On the oracle property of adaptive group Lasso in high-dimensional linear models[J]. Statistical Papers, 2016, 57(1):249-265.

[17] Huang J, Breheny P, Ma S. A Selective Review of Group Selection in High-Dimensional Models[J]. Statistical Science, 2012, 27(4):481-499.

[18] Ciuperca Gabriela. Adaptive group LASSO selection in quantile models[J]. Statistical Papers, 2019, 60:173-197.

[19] Kato K. Group Lasso for high dimensional sparse quantile regression models[J]. Statistics, 2011.

[20] Ciuperca Gabriela. Model selection by LASSO methods in a change-point model[J]. Statistical Papers, 2014, 55(2):349-374.

[21] Zhang B, Geng J, Lai L. Change-point estimation in high dimensional linear regression models via sparse group Lasso[C]// 2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton). IEEE, 2015.

[22] Ciuperca Gabriela, Matus M. Change-point detection in a linear model by adaptive fused quantile method[J]. 2019.

[23] Ciuperca Gabriela. Quantile regression in high-dimension with breaking[J]. Journal of Statistical Theory & Applications, 2013,12(3):288-305.

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

[25] Nardi Y, Rinaldo A. The log-linear group-lasso estimator and its asymptotic properties[J]. Bernoulli, 2012, 18(3):945-974.

[26] 李子强, 田茂再, 罗幼喜. 面板数据的自适应 Lasso 分位回归方法研究[J]. 统计与信息论坛, 2014, 29(07):3-10.

[27] 牟建波. 基于分位数回归的自适应组 Lasso 变量选择[D]. 西南交通大学, 2017.

[28] 牛银菊. 部分线性模型的 adaptive group lasso 变量选择[J]. 西北师范大学学报（自然科学版）, 2014(1):27-31.

[29] 崔琨鹏, 赵强. Logistic 模型自适应组 Lasso 算法[J]. 山东师范大学学报(自然科学版), 2018, 33(04):22-26.

[30] 赖学方. 基于贝叶斯统计的高维数据分析及应用[D]. 西安工程大学, 2017.

[31] 李翰芳, 罗幼喜, 田茂再. 面板数据的贝叶斯 Lasso 分位回归方法[J]. 数量经济技术经济研究, 2013, 30(02):138-149.

[32] 牟建波, 刘赫. 基于惩罚方法的贝叶斯群组变量选择[J]. 绵阳师范学院学报,

2017, 36(02):6-13.

[33] 龙振环, 张飞鹏, 周小英. 带多个变点的逐段连续线性分位数回归模型及应用[J]. 数量经济技术经济研究, 2017(08):151-162.

[34] 周小英. 逐段连续线性分位数回归模型的统计推断及其应用[D]. 湖南大学. 2018.

[35] 张立文, 倪中新, 何勇. 删失分位数回归模型中的变点检测问题[J]. 中国科学:数学, 2018, v.48(09):67-88.

[36] 郭婧, 何幼桦. 分位数自回归变点模型的贝叶斯分析及应用[J]. 统计与决策, 2017:18.

[37] Huang J, Ma S, Zhang C H . Adaptive LASSO for sparse high-dimensional regression[J]. Statistica Sinica, 2006, 18(4).

[38] Peng H, Fan J. Nonconcave penalized likelihood with a diverging number of parameters[J]. The Annals of Statistics, 2004, 32(3):928-961.

[39] He X, Shao Q M. On Parameters of Increasing Dimensions[J]. Journal of Multivariate Analysis, 2000, 73(1):120-135.

[40] Gu Y, Fan J , Kong L , et al. ADMM for High-Dimensional Sparse Penalized Quantile Regression[J]. Technometrics, 2017.

[41] Fan J, Xue L, Zou H. Strong oracle optimality of folded concave penalized estimation[J]. The Annals of Statistics, 2014, 42(3):819-849.

致谢

此时，意味着我的硕士论文即将完成，这同时也意味着，我的研究生生活即将结束。回顾这三年时光，我对身边所有的老师，朋友和同学都充满了感激之情和满满的不舍。从论文最初的选题到目前的基本完成，都离不开身边老师的悉心教导，和朋友同学的帮助于鼓励。在此，请接受我真诚的致谢，你们是我这三年中不可或缺的精彩，因为你们，我的研究生生涯更加的美丽。

本论文能够顺利完成，在此，我要特别感谢我的导师田茂再老师，在他悉心指导以及严格要求下已经完成初稿。从最初的课题选择、查找资料、大量阅读文献到文章数据模拟及实证分析的过程，无不凝结着我和导师的努力心血和辛勤汗水。在这三年的生活学习，我始终能够感受到导师对我的悉心教导和无微不至的关怀，这让我受益颇深。在此，我想借此机会表示我对导师的感激之情，对导师在照顾学生生活以及教学工作上的无私奉献精神方面充满了崇高敬意。在此期间，论文写作过程可以说是困难重重，不过在导师的帮助与指导下，在师门师兄师姐认证引导下，终于突破重重困难，文章得以顺利完成，在此过程中我也学到了许多知识，相信这段时间的学习研究对于我未来的学习及工作生活带来了极大的帮助。这段时间让我感触最深的，就是在导师的督促下以及师姐的帮助下，我从一个学术小白，慢慢懂的如何写论文，如何去研究问题。除此之外，研究生期间，最让我难以忘却的美好回忆便是我们美好而短暂的师门活动，同门师姐师弟师妹们在导师的带领下领略了大西北甘肃的自然美景，这里有其独特魅力和荒芜。在这三年时间里，我们互相鼓励互相学习，每当我们陷入矛盾困惑时，我们相约一起爬山，一起游黄河，一起骑车踏青，不仅缓解了紧张的学习压力，而且锻炼身体。

除此之外，本论文的完成与我身边的任课老师们的责任心和责任感密切相关。正是由于他们的严格要求和认真的教学，我才能够掌握和运用我学到的专业知识，在本篇论文中呈现，并最终顺利完成这篇毕业论文。在此，我要对兰州财经大学统计学院的所有老师表示衷心的感谢，并非常感谢他们过去三年来对我的淳淳教诲。