摘要

近年来，变点检测在统计学中成为一个热门方向。它在医学、工程学和金融领域等广泛应用。然而，现有的变点检测方法复杂多样，但其在检测靠近序列端点的变点时效果不理想。因此，本文研究了提高端点附近变点检测能力的相关方法，并应用到实际数据集中。

首先，本文介绍了变点检测的研究进展，并简要介绍了局部比较法、CUSUM累计法和似然比方法等相关方法。紧接着本文介绍了自规范化方法（SN）及其在检测端点附近变点时效果较差的原因，并重点介绍了能够提升检测端点附近变点能力的自适应位置自规范化方法（LASN）的相关理论。在理论介绍的基础上，本文进一步通过模拟实验，本文得出以下发现：首先，无论哪种方法，变点前后的差异越大，越容易被识别。其次，将SN方法与似然比检验方法结合使用时，未能提高检测端点附近变点的能力，几乎无法检测到端点附近的变点。第三，重点介绍的LASN方法在检测端点附近的变点时，优于传统的SN方法和似然比方法。最后，无论是LASN方法、SN方法还是似然比方法，在检测自相关程度较高的序列时都存在检验能力降低的问题。

最后，本文将能显著提高端点附近变点检测能力的LASN方法应用于美国、英国和印度三个国家的新冠疫情数据的每日累计确诊增长率数据的变点检测中。通过LASN方法找到了数据中的变点，并解释了变点出现的相关原因。

关键词：变点检测；似然比检验；自规范化方法；新冠疫情

目录

摘要 **1**

ABSTRACT **1**

1.绪论 **1**

1.1 研究背景 1
1.2 研究意义 2
1.3 文献综述 2
1.3.1 国内研究 2
1.3.2 国外研究 4
1.3.3 文献评述 5

2.变点检测常用方法 **7**

2.1 局部比较法 7
2.2 CUSUM 方法 8
2.3 似然比方法（LR） 9
2.3.1 传统似然比方法基本理论介绍 9
2.3.2 似然比方法用于序列检测的基本理论介绍 10
2.3.3 加入自规范化器的似然比变点检测理论介绍 13

3.自规范化方法的进一步研究 **16**

3.1 利用自规范化方法对均值变点的检验 17
3.2 位置自适应的单变点检测方法 19
3.3 基于位置加权的单变点检测方法 23
3.4 基于 LASN 方法的多变点检测方法 24

4.变点检测方法检验效果模拟 **25**

4.1 模型设定 25
4.2 似然比检验统计量的模拟实验 26
4.2.1 似然比检验统计量的经验接受率模拟 26
4.2.2 似然比检验统计量的经验检验功效 28
4.2.3 似然比检验统计量变点监测的检验延迟模拟 33
4.2.4 加入自规范化方法的似然比检验统计量的经验检验功效 34
4.3 SN 及相关方法的模拟实验 37

4.3.1 SN 相关方法的经验接受率和检验功效的模拟 37
4.3.2 SN 相关方法在检测端点附近变点的检验功效的模拟 39
4.3.3 LASN 方法在检测端点附近单变点的平均绝对误差 43
4.4 多变点检测的模拟实验 44
5.变点检测方法实例应用 48
5.1 对美国累计确诊人数日增长率变点的识别及成因分析 50
5.2 对英国累计确诊人数日增长率变点的识别及成因分析 52
5.3 对印度累计确诊人数日增长率变点的识别及成因分析 54
6.总结与结论 57
6.1 文章总结 57
6.2 研究结论 57
参考文献 60
致谢 63

1.绪论

1.1 研究背景

近年来，随着统计学的不断发展，变点检测问题成为了研究的热点。变点的统计推断问题主要涉及抽样理论、假设检验理论、概率极限理论等，是统计学中非常重要的研究分支。变点问题可以简单地描述如下：观察一个按时间顺序发生的随机过程，讨论在其随机变量的分布或分布参数是否有变化，或确认所观察到的随机过程是同质还是异质。同样的，在不同的研究背景下对变点检测的定义又有所区别。从变点检测个数的角度出发：变点检测可分为单变点检测和多变点检测。从变点检测的数字特征来看，变点检测又可以分成对均值、方差等等参数的检测。从检测时所考虑的数据角度来看又可将变点检测区分为在线检测和离线检测等等。

从历史上看，变点问题最初应用于质量工程管理，一般认为其研究要追溯到 1954 年 Page 在国际知名的统计杂志 Biometrika 上发表的第一篇有关变点统计分析的文章[1]。它提出的最初目的是利用连续抽样和累计和的方法对产品进行质量管理，比如一个公司或企业要对其生产线上的产品进行抽样检验质量是否合格时，需要通过相关的残次品检验来找到变点也即发生残次品的点。

目前，变点检测研究已广泛应用于金融、医学、工业等多个领域。例如在流行病学中，比较重要的问题之一是传染病在其传染过程中感染人数和累计死亡人数的变点的检测，这对确定合理的管控政策以及应对方案具有重大的意义[2]。在经济金融领域，一些重大的政治事件、卫生安全问题的发生往往会导致股价发生暴涨或暴跌[3]。工程领域的过程控制[4]、医学中的脑电图分析[5]、DNA 分割[6]等学科都有变点检测方法的具体应用。

1.2 研究意义

在实际生活中，时间序列数据十分常见。而在通过构建统计量去检测时间序列数据的变点时，由于数据具有自相关关系，往往会遇到长期方差的估计。在实际操作过程中，估计样本的长期方差需要一个基于数据本身的带宽参数，带宽参数的选择关系到假设检验的检验功效问题，如果选择的带宽参数不合适，可能会导致假设检验的功效的非单调变化。为了避免求解长期方差时的带宽选择，Shao(2010)[7]提出了的 self-normalization 的方法(下文统称 SN 方法)去检测时间序列中的变点，即通过使用自规范化式子去代替长期方差的估计，以避免非单调检验功效的产生[7]。

但是运用 SN 方法来检验靠近序列两端的变点时，检验的功效相比于检测中间位置的变点时是有所降低的。最终根据 Dai(2021)[8]的研究成果，选择运用其基于 SN 方法提出的位置自适应变点检测方法(下文统称 LASN 方法)。其方法的基本思想是基于样本点在序列中的位置，重新划分数据进行统计量的计算[8]。通过运用 LASN 方法，解决了检测靠近端点附近变点的检验功效降低的问题，也使得 SN 方法能够更加广泛地应用于需要及时发现变点的领域，比如传染病的早期变点检测以及金融等对于检测的及时性较为敏感的领域。

1.3 文献综述

1.3.1 国内研究

变点问题的国内统计研究，我国统计学者在这方面的研究始于 20 世纪 80 年代，陈希孺(1991)首先研究了常用的变点分析理论和单变点模型，阐述了常见的变点检测方法[9]。

对于变点检测的似然比类型（LR）的方法，杨喜寿和杨洪昌(1996)提出了基于 U 统计量的似然比变点检验方法，并结合了气温序列数据进行相关方法应用，通过检测气温序列数据变点，实现了气候阶段的划分[10]。赵俊(2012)则在数据呈泊松分布的假定下，提出了一种基于广义似然比的 Poisson 过程变点识别方法，同时文章也通过对应的模拟实验给出了其方法在检测变点时

的性能以及可靠性[11]。秦瑞兵、杨晓清和陈占寿等人(2019)在似然比方法的基础之上，进一步研究发现不同位置处的变点用自规范化方法进行检验时可能存在检验功效的差异，并重新提出一种基于似然比检验的变点检测方法[12]。李艳鹏(2019)则针对一元正态分布序列提出了一类含不同变点类型的多变点模型，构建了对应的似然比检验统计量，并且将其与二元分割方法相结合，将其方法由单变点检测进一步拓展到了到多变点检测[13]。祁玥(2021)则在文中将多维 Bernstein 多项式与似然比方法,CUSUM 方法以及最小二乘方法相结合，并将其运用于上证指数股票数据以及新冠肺炎数据，最终得到了较好的实验结果[14]。王丹和皮林(2021)则通过经验似然思想建立假设检验的方法，研究了重尾序列均值变点的检测问题，并且也通过模拟实验验证其方法在检验重尾序列均值变点的有效性[15]。

对于变点检测的累计和方法而言，谭常春和江敏(2020)则研究了 CUSUM 类型统计量中的调节参数的取值对变点估计结果的影响，发现不同大小调节参数的 CUSUM 类型统计量对于识别变化强度较小的变点是有显著区别的[16]。徐小平、刘君和李拂晓(2021)则基于拟极大似然估计量和 CUSUM 类型估计量，研究了面板数据中的方差变点问题，并且通过人民币汇率数据证明了两种方法在检验面板数据中的方差变点的有效性[17]。朱慧敏和王梓楠等人(2022)则基于混合误差序列，研究了方差变点模型 CUSUM 型估计量，并且将该方法运用于半导体行业的股票收益率的变点检测中，最终找到了波动率发生的位置[18]。

对于国内其他变点检测的研究方法而言，曹杰、陶云和田永丽(2002)将半截多项式引入到变点检测当中去，提出一种新的变点检测方法去对北半球天气数据进行了一个变点检测[19]。缪柏其和赵林城（2003）的研究将单变点研究逐渐延伸到研究多变点问题变点个数和变点位置的检测，并且利用极大似然检测方法检测变点并对金融传染进行分析[20]。金融市场中，隋学深和杨忠海(2007)运用贝叶斯变点检测方法，实现了对上证综合指数月度时间序列数据的变点检测[21]。李傲梅(2019)则基于 GMD 算法中的自适应组 Lasso 方法去估计多元线性回归模型中的变点，同时将该方法应用于和空气质量有关的实际数据中[22]。

1.3.2 国外研究

而对于变点检测问题的国外统计研究，Page(1954)被广泛认为是第一次引入了未知的变点问题的相关文献。其最初目的是利用连续抽样和累计和的方法对生产线上的产品进行一个质量管理[23]。随后的几十年中，各种变点检测文章都纷纷涌现出来。对于综述类型的文献而言，Truong(2020)在文中向介绍了变点的定义以及不同的离线变点检测方法和对应的评价指标[24]。

对于变点检测的似然比类型的方法，Quandt 和 Laurent 等人(1958)将变点监测的类型扩展到了线性模型，并且首次提出了解决该问题的似然比检验方法（LR）[25]。在之后的研究中，Yao 和 Davis 等人(1984)则研究了服从独立同正态分布数据中均值变点的 LR 检验的渐近性质。并且发现当序列中变点的位置位于序列中部时，贝叶斯检验方法是要优于似然比 LR 检验方法的[26]。进一步，对于多变点问题，Bai 和 Perron(1998)进一步提出了一种回归模型中多变点 LR 检验方法，用于确定变点位置和数量[27]。之后，Holger 和 Gösmann(2020)将 LR 检验方法运用到高维的时间序列变点检测，并且将其与自规范化方法相结合，展示了其方法在检验均值变点，方差变点以及中位数变点等相关统计量时的效果，最后还将相关方法运用到了互联网泡沫指数价格数据中[28]。Dette 和 Gösmann(2022)则进一步将似然比类型的检验方法运用到存在时间和空间的依赖性的高维水文数据中，验证了其对高维有依赖性数据的变点检测效果是优于 CUSUM 方法的[29]。

而对于变点检测的累计和类型的方法，Brown 和 Durbin(1975)等提出了累计平方和方法讨论时间序列中的方差变点问题，并给出了方差变点的渐近分布[30]。Carla 和 Inclán 等人(1994)在 CUSUM 算法的基础上进行优化，提出了迭代累加平方和算法，之后该方法便被广泛应用于金融时间序列中去[31]。之后的时间中 Kokoszka 和 Leipus 等人(1998)又进一步证明了 CUSUM 统计量的一致性，并且利用 CUSUM 方法完成了 ARCH 模型中的均值变点检测，得到了较好的结果[32]。随后，Lee 和 Maekawa(2004)等人在对残差累积平方和假设检验的基础上，分析了 GARCH(1，1)模型的参数变化，给出了 CUSUM 检验统计量的极限分布，并将其应用于美元汇率数据的变点检测[33]。

但是无论是LR类型的检验统计量还是CUSUM类型的检验统计量，当处理时间序列类型的数据的时候，往往需要计算一致的长期方差估计量，且长期方差估计量的计算还涉及到一个带宽参数的选择。为了解决这一难题，Shao(2010)提出了用一个自规范化器代替长期方差的估计方法，使其能以统一的方式处理广泛类别的参数（例如均值、方差、相关性和分位数）的变点检测，并且且实现了检验功效的单调变化[7]。之后Shao(2015)进一步详细讨论了在自规范化背景下相关参数的置信区间的构造，给出了时空数据以及空间数据自规范化器的构造方法。同时还总结了自归范化在时间序列背景下的优点以及局限性[34]。在自规范化方法的基础之上，Lavitas和Zhang等人(2018)进一步提出一种非监督变点检测方法，并通过蒙特卡罗模拟证明了其方法的有限样本性质[35]。Jiang和Zhao等人(2021)在自规范化方法的基础上结合了Narrowest-Over-Threshold (NOT)算法[36]，进一步拓展了基于自规范化方法的变点检测能力，且将基于自规范化方法检测单个变点拓展到检测多个变点问题，并且成功地运用到Covid_19累计确诊和累计死亡的时间序列数据中[2]。Dai和She(2021)通过对自规范化方法的进一步研究，发现传统SN方法在检验端点附近变点的时候有时同样会出现检验功效的非单调变化并给出其解决办法，同时在最后还给出基于其统计量的多变点检测算法，有效地拓展了自规范化方法检测变点的能力[8]。Wang和Shao等人（2021）在之前运用自规范化方法研究一维时间序列的基础之上，进一步将U统计量和自规范方法结合起来实现了对高维时间序列数据进行变点的检测[37]。

1.3.3 文献评述

自从变点问题被提出以来，已经引起了国内外许多学者的关注。随着社会的飞速发展，国内外文献的变点问题的研究已经从最初的工程方面的质量控制逐步延伸到经济金融，公共卫生，水文勘测等多个领域之中。从其发展的角度来看，变点检测从最初的单变点模型发展到了多变点模型，检测维度从一维发展到了高维，并且随着变点检测研究越来越复杂多样，其应用性也在不断增强。

从所读文献来看，变点问题已经被国内外学者大量的研究，但也有许多问题是值得我们去研究的。首先，再目前的研究中，单变点问题的研究已经发展地较为完善，但是实际的生活中会存在众多变点位置和数量未知的多变点问题，这个领域发展还不成熟。其次，许多参数模型往往会要求数据分布，但是现实生活中往往难以获取数据的具体分布情况，如何利用合理的估计形式进行有效的变点检测也是需要深入研究的。第三，目前对于变点检测众多方法，大部分检测方法在检测靠近序列中间位置的变点效率较高，而在检测靠近端点附近的变点却相对较差，检测端点附近的变点的方法还是存在发展空间的。

2.变点检测常用方法

在介绍变点检测常用到的方法之前，首先给出变点问题的相关定义：不妨假设有这样的样本序列 $X_1,.....,X_{n-1},X_n$ ，其对应的分布函数分别为 $F_1,.....F_{n-1},F_n$。如果存在时点 k^* 使得序列的某一数字特征比如均值或者方差，在 k^* 之前和在 k^* 之后有较大的差异，则称时刻 k^* 为序列 $X_1,.....,X_{n-1},X_n$ 的一个变点，有时候也常记为 $\lambda^* = k^*/n$，即变点在整个序列中的分位点来表示变点的位置。

更一般地，若对于序列 $X_1,.....,X_{n-1},X_n$ ，有这样的 j 种划分，即 $\{X_1,...,X_{k_1}\},\{X_{k_1+1},....,X_{k_2}\},...,\{X_{k_j+1},....,X_n\}$，使得每一组内样本的分布较为稳定，而在 $k_1,...,k_j$ 处有突变，则称 $k_1,k_2,...k_j$ 为序列 $X_1,.....,X_{n-1},X_n$, 的 j 个变点。

目前，学术界还未给出变点的严格的定义，对于不同的研究领域，其定义也大都有所差别，但检测变点的许多经典方法确是通用的，下面来一一介绍。

2.1 局部比较法

局部比较法的主要思想为在变点附近的某个部分序列中，感兴趣的特征发生了显著性的变化，它可以通过适当的估计量显示出来。而在非变点附近的部分序列中，其估计量保持稳定。因此针对统计量在各个局部片段是否变化，取其显著之处作为变点的估计。

首先考虑如下的待检测的单变点模型：

$$X_i = \begin{cases} \mu_1 + e_i, 1 \le i \le k^* \\ \mu_2 + e_i, k^* \le i \le n \end{cases} \tag{2.1}$$

其中随机误差 $\{e_i\}_{i=1}^n$ 为独立同分布的随机变量，且 $E(e_i)=0$，$Var(e_i)=\sigma^2<+\infty$，

μ_1， μ_2分别为变点k^*之前和之后的变化强度。

有了上述模型序列，对变点k^*的局部比较法估计步骤如下：

第一步，对每一个潜在的变点i， $i \in \{1,2,..,n-1,n\}$，给定一个区间长度l，将i左右各l个观察值求和并相减，即得：

$$\begin{aligned} &Y_i = (X_i + ... + X_{i+l-1}) - (X_{i-l} + ... + X_{i-1}) \\ &i = l+1, l+2, ..., n-l+1 \end{aligned} \tag{2.2}$$

第二步，若i不是变点，则Y_i更加趋于0。反之，若i恰好是变点或者变点恰好落在i得左l邻域或者右l邻域。则Y_i左右两项之差会较大，因此有下式：

$$\left|Y_{k^*}\right| = \max(\left|Y_{l+1}\right|, \left|Y_{l+2}\right|, ..., \left|Y_{n-l+1}\right|) \tag{2.3}$$

第三步，当$\left|Y_{k^*}\right|$超过某个临界值时，则拒绝原假设，认为序列中存在变点，变点k^*即为使得$\left|Y_i\right|$取得最大值的下标。否则的话不拒绝H_0，即没有充分证据证明序列中存在均值变点。

2.2 CUSUM 方法

CUSUM 方法是处理均值变点分析问题的最常见的方法，无论是国内还是国外都有大量学者研究使用。在涉及到均值变点或者可以将变点问题转化为均值变点的研究都可以使用 CUSUM 方法。其基本思想就是对每一潜在变点前后的微小偏移加以累计，通过 CUSUM 统计量以检验是否存在异常趋势。

首先同样考虑式子(2.1)中的单变点模型：

$$X_i = \begin{cases} \mu_1 + e_i, 1 \le i \le k^* \\ \mu_2 + e_i, k^* \le i \le n \end{cases} \tag{2.4}$$

其中随机误差$\{e_i\}_{i=1}^n$为独立同分布的随机变量，且$E(e_i) = 0$， $Var(e_i) = \sigma^2 < +\infty$，$\mu_1$， μ_2分别为变点k^*之前和之后的变化强度。则变点k^*的传统的 CUSUM 估计步骤如下：

第一步：通过 CUSUM 检验统计量计算每一个潜在变点的检验统计值U_k，而其计算的公式如下：

$$U_k = (\frac{k(n-k)}{n})^{1-\gamma}\{\frac{1}{k}\sum_{i=1}^{k}X_i - \frac{1}{n-k}\sum_{i=k+1}^{n}X_i\} \quad (2.5)$$
$$0 \le \gamma \le 1$$

第二步：将计算出来的检验统计值U_k与临界值U进行对比。

第三步：若对于每一个潜在变点k，U_k的值若都小于临界值U，则没有充分证据证明序列中存在变点，反之若存在k，U_k的值大于临界值U，则变点k^*为

$$k^* = \min\{k : |U_k| = \max_{1\le j\le n}|U_j|\} \quad (2.6)$$

2.3 似然比方法（LR）

2.3.1 传统似然比方法基本理论介绍

似然比变点检测方法(LR)是变点理论研究进程中较早就提出的理论，该方法一般用于检验正态分布序列均值变点，为保持和 2.1 节和 2.2 节中序列模型的一致，下面也同样给出序列中仅存在一个变点的序列模型。

设随机变量序列$X_1,.....,X_{n-1},X_n$,之间相互独立，其分布具体形式如下：

$$X_i = \begin{cases} \mu_1 + e_i, 1 \le i \le k^* \\ \mu_2 + e_i, k^* \le i \le n \end{cases} \quad (2.7)$$

其中随机误差$\{e_i\}_{i=1}^n$为独立同分布的随机变量，且$E(e_i)=\mu$，$Var(e_i)=\sigma^2<+\infty$，$\mu_1$，$\mu_2$分别为变点$k^*$之前和之后的变化强度。有了假定的序列，利用似然比变点检测方法(LR)检测变点的步骤如下：

第一步：对每一个潜在的变点k，$k\in\{2,..,n-1,n\}$，计算每一个位置的似然比检验统计值：

$$ML_k = \frac{\sup_{(\mu,\sigma^2)\in\Theta_0}\prod_{i=1}^{n}f(x_i;\mu,\sigma^2)}{\sup_{(\mu_1,\mu_2,\sigma^2)\in\Theta}\prod_{i=1}^{k}f(x_i;\mu_1,\sigma^2)\prod_{i=k+1}^{n}f(x_i;\mu_2,\sigma^2)} \quad (2.8)$$
$$G_k = -2\ln ML_k$$

在 ML_k 中，$f(\cdot)$ 待为假定序列的密度函数，μ 为整个待检测序列的均值，μ_1，μ_2 分别为潜在变点 k 前后的样本均值，σ^2 代表整个序列的样本方差，假定其是不变的。

第二步：对比检验统计值和临界值大小，不妨取检验统计量为 $G_k = \max_{2\le k\le n}(-2\ln ML_k)$。

第三步：找变点位置，若检验统计量的值小于临界值，则认为序列中不存在均值变点，否则的话序列中存在均值变点，此时估计的变点为：

$$k^* = \arg\{k : G = \max_{2\le k\le n} G_k\} \tag{2.9}$$

2.3.2 似然比方法用于序列检测的基本理论介绍

通过对上一小节传统的似然比检验方法的分析阐释，对于变点检测而言，传统的似然比方法都是专注于后验变化点的分析。即对于一个完整的历史数据集去研究其结构化变点问题。而另一方面，对于工程学，医学，金融学等学科领域，其收集到的数据都是稳定的增加的，及时识别新搜集到的变点对于构建模型以及进行政策指定和预警。紧接着似然比方法理论，在下面的部分来进一步说明利用似然比方法来进行对序列的监测问题。

首先不妨假定已经有一段长度为 m 的原始平稳的时间序列，$X_1,\ldots,X_m$，在后期继续观测到 k 个新增序列，$X_{m+1},\ldots,X_{m+k}$。Claudia[38]其通过直接对比前 m 项平稳序列和后 k 项序列均值之差来达到构造统计量的效果，也就类似于下文的统计量 Q。而 Dette[28]其通过对比 $X_1,\ldots,X_{m+j}$ 与 $X_{m+j+1},\ldots,X_{m+k}$ 均值之差来构建对应的检验统计量，也就类似于下文的统计量 P。Dette 的方法相对于 Claudia 的方法而言，是能够有效地提升变点的检测能力。

下面通过假设检验的方法来详细说明似然比检验统计量如何对一维时间序列均值变点做检验，首先设定所研究的时间序列为 $\{X_t\}_{t\in\mathbb{Z}}$，感兴趣的是希望了解在时间点 $m+k$，序列的均值相对于最初那一段平稳序列是否发生了变化，其中 $m+k\ge m+1$。因此建立如下的原假设与备择假设：

$$H_0 : \mu_1 = \cdots = \mu_m = \mu_{m+1} = \cdots = \mu_{m+k},$$

$$H_1 : \exists j \in \{0,\ldots,k-1\} : \mu_1 = \cdots = \mu_{m+j} \neq \mu_{m+j+1} = \cdots = \mu_{m+k}$$

若原假设 H_0 成立，则代表着原序列中待检测部分 $X_{m+1},\ldots,X_{m+k}$ 中未发现均值发生变化的时间点。而若原假设 H_0 被拒绝，则说明序列中至少存在一个点发生了均值的突变。

有了原序列与待检测序列 $X_1,\ldots,X_{m+k}$，根据可以依照 Dette(2020)中的似然比的思想去构造在潜在变点处的似然统计量：

$$\Lambda_m(k) = \frac{\sup_{\mu\in\mathbb{R}^d} \prod_{t=1}^{m+k} f\left(X_t,\mu\right)}{\sup\limits_{\substack{j\in\{0,\ldots,k-1\}\\ \mu^{(1)},\mu^{(2)}\in\mathbb{R}}} \prod_{t=1}^{m+j} f\left(X_t,\mu^{(1)}\right)\cdot \prod_{t=m+j+1}^{m+k} f\left(X_t,\mu^{(2)}\right)} \tag{2.10}$$

其中 $f(\cdot,\mu)$ 代表在对应随机变量处的概率密度函数，μ 代表时间点 $m+k$ 前的样本均值，$\mu^{(1)}$ 代表潜在变点前的样本均值，$\mu^{(2)}$ 代表潜在变点后的样本均值。为了简化式(2.5)的计算，对式(2.5)做取对数处理。

$$\begin{aligned} -2\log\left(\Lambda_m(k)\right) &= \max_{j=0}^{k-1} \frac{(m+j)(k-j)}{m+k}\left(\hat{\mu}_1^{m+j} - \hat{\mu}_{m+j+1}^{m+k}\right)^{\top} \times \Sigma^{-1}\left(\hat{\mu}_1^{m+j} - \hat{\mu}_{m+j+1}^{m+k}\right) \\ &= \max_{j=0}^{k-1} \frac{(m+k)(m+j)}{(k-j)}\left(\hat{\mu}_1^{m+j} - \hat{\mu}_1^{m+k}\right)^{\top} \times \Sigma^{-1}\left(\hat{\mu}_1^{m+j} - \hat{\mu}_1^{m+k}\right) \end{aligned} \tag{2.11}$$

对于简化后的式子(2.5)，$^{\top}$ 表示对向量的的转置。Σ 代表的是长期方差，常用最初的平稳段时间序列数据加以计算，在后文中也会给出其相应的计算公式。

$$\hat{\mu}_i^j = \frac{1}{j-i+1}\sum_{t=i}^{j} X_t$$

有了上面对似然比统计量的简化，可以认为当统计量(2.12)的值大于对应的临界值的时候，往往有把握拒绝原假设认为在新添加的序列中至少存在一个均值变点。

$$\max_{j=0}^{k-1}(m+j)(k-j)\left(\hat{\mu}_1^{m+j} - \hat{\mu}_{m+j+1}^{m+k}\right)^{\top} \Sigma^{-1}\left(\hat{\mu}_1^{m+j} - \hat{\mu}_{m+j+1}^{m+k}\right) \tag{2.12}$$

对于(2.12)式似然比类型的统计量而言，其渐近性质较难研究，最终其建议选择带有加权类型的似然比统计量，其具体形式如(2.13)所示：

$$\hat{D}_m(k)=m^{-3}\max_{j=0}^{k-1}(m+j)^2(k-j)^2\left(\hat{\mu}_1^{m+j}-\hat{\mu}_{m+j+1}^{m+k}\right)^{\top}\times\hat{\Sigma}_m^{-1}\left(\hat{\mu}_1^{m+j}-\hat{\mu}_{m+j+1}^{m+k}\right) \quad \textbf{(2.13)}$$

与此同时，为了对比似然比类型的统计量同其他类型检验统计量在检测变点类型时的能力强弱，除了统计量 $\hat{D}_m(k)$ 外，式子(2.14)，(2.15)还加入另外两种类似的统计量，一种是

$$\hat{Q}_m(k)=\frac{k^2}{m}\left(\hat{\mu}_1^{m}-\hat{\mu}_{m+1}^{m+k}\right)^{\top}\hat{\Sigma}_m^{-1}\left(\hat{\mu}_1^{m}-\hat{\mu}_{m+1}^{m+k}\right) \quad \textbf{(2.14)}$$

另一种是

$$\hat{P}_m(k)=\max_{j=0}^{k-1}\frac{(k-j)^2}{m}\left(\hat{\mu}_1^{m}-\hat{\mu}_{m+j+1}^{m+k}\right)^{\top}\hat{\Sigma}_m^{-1}\left(\hat{\mu}_1^{m}-\hat{\mu}_{m+j+1}^{m+k}\right) \quad \textbf{(2.15)}$$

其中 $\hat{D}_m(k)$ 和 $\hat{Q}_m(k)$ 以及 $\hat{P}_m(k)$ 的主要差别是 $\hat{D}_m(k)$ 在计算对比统计量时，其在计算检验统计量时的样本选取是有所差异的。$\hat{Q}_m(k)$ 以及 $\hat{P}_m(k)$ 两种方法都选择原始序列长度 m 的样本均值与 m 之后的每一个潜在变点进行对比。而似然比检验统计量则会选取潜在变点 k 之前的所有的样本均值作为对比基础，并不会像统计量 $\hat{Q}_m(k)$ 以及 $\hat{P}_m(k)$ 一样对比统计量仅仅选择初始序列长度 m。对于长期方差 Σ，由于目前所研究的是一维的序列，因此样本的长期方差选用式(2.16)来对其进行估计。

$$\hat{\sigma}^2=\hat{\gamma}_0+2\sum_{i=1}^{m-1}k\left(\frac{i}{b_m}\right)\hat{\gamma}_i \quad \textbf{(2.16)}$$

其中 $\hat{\gamma}_i$ 代表的是原始平稳序列 $X_1,\ldots,X_m$ 的滞后 i 阶的经验自协方差，b_m 代表的是计算时所采取的平滑带宽。$k(x)$ 的具体形式如下。

$$k(x)=\frac{25}{12\pi^2x^2}\left(\frac{\sin(6\pi x/5)}{6\pi x/5}-\cos(6\pi x/5)\right) \quad \textbf{(2.17)}$$

有了上述的三种检验统计量的计算公式，可以比较容易地计算出每一个潜在变点位置 k 的检验统计值。当潜在变点位置 k 的检验统计值大于对应的临界值的时候，便可以有信心拒绝对应显著性水平下的原假设，认为该点可能是一个均值变点。下面给出三种统计量对应的临界值的计算方法。

对于统计量 $\hat{D}_m(k)$ 而言，要去寻找常数 C_α，当 $\hat{D}_m(k)>C_\alpha w(k/m)$ 时，便可以拒绝对应显著性水平下的原假设，认为序列中存在变点。即 C_α 应该满足

$$\limsup_{m\to\infty}\mathbb{P}_{H_0}\left(\max_{k=1}^{mT}\frac{\hat{D}_m(k)}{w(k/m)}>C_\alpha\right)\le\alpha \tag{2.18}$$

其中 $w(.)$ 表示选取的对应的权重函数，后文会进一步详细说明选取形式，而最终所得到的潜在变点位置 τ_m 可以如下表示：

$$\tau_m=\inf\left\{1\le k\le \mathrm{T^*m}\mid \hat{D}_m(k)>C_\alpha w(k/m)\right\} \tag{2.19}$$

为了进一步通过蒙特卡洛模拟实验来计算其对应显著性水平下的临界值，相关的学者同样给出了三种检验统计量的渐近分布，其具体形式如下。

对于统计量 $\hat{D}_m(k)$ 而言，当 m 趋于无穷的时候：

$$\max_{k=1}^{\mathrm{Tm}}\frac{\hat{D}_m(k)}{w(k/m)}\overset{\mathcal{D}}{\Rightarrow}\sup_{t\in[1,T+1]}\sup_{s\in[1,t]}\frac{B(s,t)^\top B(s,t)}{w(t-1)} \tag{2.20}$$

对于统计量 $\hat{Q}_m(k)$ 而言，当 m 趋于无穷的时候：

$$\max_{k=1}^{\mathrm{Tm}}\frac{\hat{Q}_m(k)}{w(k/m)}\overset{\mathcal{D}}{\Rightarrow}\sup_{t\in[1,T+1]}\frac{B(t,1)^\top B(t,1)}{w(t-1)} \tag{2.21}$$

对于统计量 $\hat{P}_m(k)$ 而言，当 m 趋于无穷的时候：

$$\max_{k=1}^{Tm}\frac{\hat{P}_m(k)}{w(k/m)}\overset{\mathcal{D}}{\Rightarrow}\sup_{t\in[1,T+1]}\sup_{s\in[1,t]}\frac{(B(1,s)+B(t,1))^\top(B(1,s)+B(t,1))}{w(t-1)} \tag{2.22}$$

其中

$$B(s,t)=tW(s)-sW(t) \tag{2.23}$$

$\{W(s)\}_{s\in[0,T+1]}$ 代表区间 $[0,T+1]$ 上的布朗运动。

2.3.3 加入自规范化器的似然比变点检测理论介绍

通过 2.3.2 的介绍，无论是似然比统计量 $\hat{D}_m(k)$ 还是 $\hat{Q}_m(k)$，还是 $\hat{P}_m(k)$，在对其检验统计量进行计算的过程中，不可避免地需要根据原始序列 $X_1,\ldots,X_m$ 去计算长期方差 Σ。计算长期方差 Σ 涉及到人为地去选择平滑带宽 b_m，若 b_m 选的不是十分恰当，相关学者通过研究发现可能会造成变点检测的效果较差以至于难以准确识别变点，因此在原统计量中加入一个自规范化器

去避免对长期方差的估计从而也避免了人为地去选择自规范化器而导致的检验功效地非单调变化。因此，前文的似然比检验统计量加入自规范化器 $\mathbb{V}(.)$ 之后的统计量形式为

$$\begin{aligned}\hat{D}^{\mathrm{SN}}(k)=m\max_{j=0}^{k-1}(m+j)^2(k-j)^2\left(\hat{\mu}_1^{m+j}-\hat{\mu}_{m+j+1}^{m+k}\right)^{\top}\\ \times\mathbb{V}^{-1}(m+j,m+k)\left(\hat{\mu}_1^{m+j}-\hat{\mu}_{m+j+1}^{m+k}\right)\end{aligned} \tag{2.24}$$

其中自规范化器 $\mathbb{V}(.)$ 的具体形式如下：

$$\begin{aligned}\mathbb{V}(z,u)=&\sum_{j=1}^{z}j^2(z-j)^2\left(\hat{\mu}_1^{j}-\hat{\mu}_{j+1}^{z}\right)\left(\hat{\mu}_1^{j}-\hat{\mu}_{j+1}^{z}\right)^{\top}\\ &+\sum_{j=z+1}^{u}(u-j)^2(j-z)^2\left(\hat{\mu}_{z+1}^{j}-\hat{\mu}_{j+1}^{u}\right)\left(\hat{\mu}_{z+1}^{j}-\hat{\mu}_{j+1}^{u}\right)^{\top}\end{aligned} \tag{2.25}$$

同理，对于统计量 $\hat{P}_m(k)$ 而言，由于其检验统计量的计算过程中同样需要使用到长期方差，同样加入自规范化其去代替长期方差的估计，最终加入自规范化器的统计量如下：

$$\begin{aligned}\hat{P}_m^{\mathrm{SN}}(k)=m^3\max_{j=0}^{k-1}(k-j)^2\left(\hat{\mu}_1^{m}-\hat{\mu}_{m+j+1}^{m+k}\right)^{\top}\\ \times\mathbb{V}^{-1}(m+j,m+k)\left(\hat{\mu}_1^{m}-\hat{\mu}_{m+j+1}^{m+k}\right)\end{aligned} \tag{2.26}$$

当有了基于自规范化方法的检验统计量后，可以较为容易地去计算在原始序列 $X_1,\ldots,X_m$ 之后的各个潜在变点位置的检验统计量，从而通过对比对应显著性水平的临界值来判断变点发生的时刻。由于加入了自规范化器，在计算自规范化类型的统计量的临界值的蒙特卡洛模拟也会发生变化，相关的学者同样也给出了所需要的统计量的渐近分布，其具体形式如下。

$$\max_{k=1}^{\mathrm{Tm}}\frac{\hat{D}_m^{\mathrm{SN}}(k)}{w(k/m)}\overset{\mathcal{D}}{\Rightarrow}\sup_{t\in[1,T+1]}\sup_{s\in[1,t]}\frac{|\tilde{B}(s,t)|}{w(t-1)} \tag{2.27}$$

其中

$$\tilde{B}(s,t)=B^{\top}(s,t)\left(N_1(s)+N_2(s,t)\right)^{-1}B(s,t)$$

$$N_1(s)=\int_0^s B(r,s)B^{\top}(r,s)dr$$

$$N_2(s,t)=\int_s^t\left(B(r,t)+B(s,r)-B(s,t)\right)\left(B(r,t)+B(s,r)-B(s,t)\right)^{\top}dr$$

而对于统计量 $\hat{P}_m^{\mathrm{SN}}(k)$，其由于同样使用了自规范化其去代替长期方差的估计，其统计量的渐近分布同样发生了变化，其渐近分布可以通过如下式子加以表示

$$\max_{k=1}^{Tm} \frac{\hat{P}_m^{\mathrm{SN}}(k)}{w(k/m)} \Rightarrow^{\mathcal{D}} \sup_{t\in[1,T+1]} \sup_{s\in[1,t]} \frac{\left|\tilde{B}^{\mathrm{SN}}(s,t)\right|}{w(t-1)} \tag{2.28}$$

其中：

$$\tilde{B}^{\mathrm{SN}}(s,t) = (B(1,s)+B(t,1))^{\top}\left(N_1(s)+N_2(s,t)\right)^{-1}$$
$$(B(1,s)+B(t,1))$$

但是，值得注意的是，虽然统计量 $\hat{D}_m(k)$，$\hat{P}_m(k)$ 能够通过加入自规范器的方式避免对长期方差进行估计，但是统计量 $\hat{Q}_m(k)$ 是无法加入自规范器来避免对长期方差进行估计。原因是无论是 $\hat{D}_m(k)$ 还是 $\hat{P}_m(k)$，在计算统计量的时候都会用到在变点前和变点后分离计算最大对比统计量的思想，而 $\hat{Q}_m(k)$ 在构建统计量时直接计算的是当前潜在变点的统计量值，未用到求最大对比统计量的思想，因此无法对其进行自规范化。而关于变点的自规范化方法在第三章中具体来介绍。

3.自规范化方法的进一步研究

通过对第 2 章以及过往文献的研究发现，人们往往会假定某些参数化模型的具体形式，尤其是第二章中的似然比类型的序列监测检验统计量，其主要检验的数据类型是正态分布。除此之外，检验那些独立同分布的一些随机变量是否存在变点，并且当研究的随机变量为时间序列数据时，还应该将其序列相关性纳入到检测程序中去。而序列的相关性由序列的长期方差反映，即随机变量序列的所有阶数的自协方差之和来反映。而正如第 2 章对似然比类型的检验统计量相关理论的阐述，对长期方差的估计，如式子(2.16)，还往往会涉及到一个至关重要的带宽 b_m 的选择，带宽的选择直接关系到假设检验的统计功效，从以往的文献 Dette[28]上来看，一个不恰当的带宽选择往往会导致检验功效的非单调变化。比如说随着变化强度的增加，对变点的假设检验的检验功效反而下降的现象。

为了避免对长期方差估计时由于带宽选择的不恰当导致的非单调检验功效的出现，Shao[7]基于自规范化的思想，将长期方差用一个自规范化式子去代替，从而避免了对长期方差的估计，其方法被称为自规范化方法。模拟实验也证明基于自规范化的方法(下文用 SN 方法来指代)在对于单变点问题的假设检验中有不错的检验功效，同时也具有不错的检验水平。

尽管 SN 方法在检测变点时相较于其他基于模型假定的研究方法有众多的优势，但是 SN 方法在检测位于端点处的变点时，其检验的统计功效远远不如变点位于序列中间位置。也是在此不足的基础上，本文希望提升构造的假设检验对于端点处变点检测的能力。因此，下面将自规范化理论研究部分进一步分成以下几个小的部分来对端点处的变点进行理论研究：

第一部分将从检验有无均值变点入手，介绍传统自规范化方法，同时对该方法在端点处的检验功效降低做出说明。第二部分介绍基于 SN 方法改进的适应不同潜在变点的位置的单变点检测方法，后文统称 LASN 方法，并且给出

构造该统计量的最优样本选取方法。第三部分提出基于位置加权的变点检测方法。第四部分将提出单边点拓展到多个变点的变点检测理论，并介绍具体的变点搜索的方法和步骤。

3.1 利用自规范化方法对均值变点的检验

为了说明基本方法，首先考虑对一随机变量序列的均值是否存在变点进行研究。不妨假设 $X_1,.....,X_n$，为一系列观测到的单变量时间序列。为了对其进行序列均值的变点检测，构造如下的假设检验：

$$H_0 : E\left(X_1\right) = \ldots = E\left(X_n\right)$$
$$H_1 : E\left(X_1\right) = \ldots = E\left(X_{k^*}\right) \neq E\left(X_{k^*+1}\right) = E\left(X_n\right)$$

其中 k^* 表示变点的实际位置，且通常在实际中是无法事先知晓的。

为了方便讨论，记序列 $X_j,.....,X_k$ 的样本均值为 $\bar{X}_{j,k} = (k-j+1)^{-1}\sum_{i=j}^{k} X_i$，在众多的文献中，用的比较多的方法就是使用递归不断地将部分均值 $\bar{X}_{1,k}$ 和序列总体均值作差，从而生成一个之前所介绍的使用较为广泛的 CUSUM 统计量：

$$Z_n(\lfloor nt \rfloor) = \frac{1}{\sqrt{n}} \sum_{1}^{\lfloor nt \rfloor} \left(X_i - \bar{X}_{1,n}\right) \tag{3.1}$$

其中 $\lfloor nt \rfloor$ 表示 nt 向下取整，$t \in [0,1]$。为了说明 $Z_n(\lfloor nt \rfloor)$ 的渐近性质，还需要下面较为宽松的假定：当 $n \to \infty$ 时

$$S_n(\lfloor nt \rfloor) = \frac{1}{\sqrt{n}} \sum_{1}^{\lfloor nt \rfloor} \left\{X_i - E\left(X_i\right)\right\} \Rightarrow \sigma B(t) \tag{3.2}$$

其中“$\Rightarrow$”表示弱收敛，$B(t)$ 表示标准布朗运动且 σ^2 表示其长期方差。若(3.1)成立，根据连续映射定理在原假设成立的条件下：

$$Z_n(\lfloor nt \rfloor) = S_n(\lfloor nt \rfloor) - \frac{\lfloor nt \rfloor}{\sqrt{n}} S_n(n) \Rightarrow \sigma\{B(t) - tB(1)\} \tag{3.3}$$

然而 $\sigma\{B(t)-tB(1)\}$ 的渐近分布又依赖于长期方差中的 σ 参数，一个比较直观的做法就是用一个 σ 的一致估计量去代替，但是，这又涉及到在前文所说的对于 σ^2 的估计往往会涉及到一个平滑带宽的选择，基于数据的带宽的选择

还必须要与数据内在结构相适应，因此当原假设被拒绝时，用其他统计量去代替 σ^2 会造成检验功效的下降。

为了避免对 σ^2 的估计，Shao 提出了使用自规范化的方法去消除 σ，并且在原假设成立的条件下，得到了统计量的渐近分布：

$$G_n = \max_{k=1,\ldots n-1} Z_n^2(k)/V_n(k) \xrightarrow{d} \sup_{t\in[0,1]} \{B(t)-tB(1)\}^2/V(t) \tag{3.4}$$

其中“$\xrightarrow{d}$”表示依分布收敛，$V_n(k)$ 在具体样本序列中的计算公式如下：

$$\begin{aligned} V_n(k) &= \frac{1}{n^2}[\sum_{j=1}^{k}\{\sum_{i=1}^{j}(X_i-\overline{X}_{1,k})\}^2 + \sum_{j=k+1}^{n}\{\sum_{i=j}^{n}(X_i-\overline{X}_n)\}^2] \\ V(t) &= \int_0^t \{B(s)-\frac{s}{t}B(t)\}^2 ds + \int_t^1 \{B(1)-B(s)-\frac{1-s}{1-t}[B(1)-B(t)]\}^2 ds \end{aligned} \tag{3.5}$$

当拒绝原假设时，即序列中存在变点时，检验统计量 G_n 依概率收敛到无穷，从而使得假设检验的功效值接近于 1。然而，对于那些距离端点处非常近即，k^* 接近于 1 或者 n 时，通过之后的的模拟实验很容易发现，相比于 k^* 位于中间位置，检验端点附近潜在的变点时，利用 SN 方法做假设检验的统计功效会大幅下降，下面来说明原因。

假定 $k^* = \lfloor n\lambda^* \rfloor$，其中 $\lambda^* \in (0,1)$，λ^* 趋于 0 或者 n 时，意味着变点处于起始端点或者末端。当原假设被拒绝时，$Z_n\left(k^*\right)$ 可以分解为如下形式

$$Z_n(k^*) = S_n(k^*) - \frac{k^*}{\sqrt{n}}S_n(n) - \frac{k^*}{\sqrt{n}}(1-\frac{k^*}{n})\Delta_n \tag{3.6}$$

其中 Δ_n 代表随机变量序列均值变化的幅度，并且当 $|\Delta_n|\sqrt{n} \to \infty$ 时，在式(3.6)的条件下，$Z_n\left(k^*\right)$ 可以进一步写为如下形式：

$$Z_n\left(k^*\right) \sim \sqrt{n}\lambda^*\left(1-\lambda^*\right)\Delta_n \tag{3.7}$$

其中“~”表示 $Z_n\left(k^*\right)$ 渐近等价于 $\sqrt{n}\lambda^*\left(1-\lambda^*\right)\Delta_n$。

通过上式，发现由于检验统计量 G_n 的分母部分 $V_n(k^*)$ 的概率测度最大数量级不超过 1，且其值不依赖于 Δ_n，因此 G_n 的大小主要是取决于 $\lambda^*\left(1-\lambda^*\right)$ 以

及Δ_n。而由式$\lambda^*\left(1-\lambda^*\right)$，当变点趋近于两端时即$\lambda^*$趋于 1 或者 0 时，$G_n$的值会变得非常小，因此会会很难拒绝原假设，从而很难发现端点附近的变点，这也是为何 SN 方法在端点处检测变点时其检验功效很低的原因。

3.2 位置自适应的单变点检测方法

为了解决端点处传统 SN 方法检验功效的下降，DAI[8]提出了在原 SN 方法的基础上，依据潜在变点位置重新选择计算检验统计量的样本，并通过模拟发现新的检验统计量计算方法能够提升 SN 方法检测端点附近的变点。

通过式子(3.1)和(3.5)，发现传统的 SN 方法在计算检验统计量时，无论潜在变点位置位于序列中间位置还是位于靠近序列的两端，传统的 SN 方法每一次计算检验统计量时都会对比潜在变点整个左边部分均值和整个右边部分均值。而这就造成了一个问题，比如当计算靠近左端点处的变点的对应的统计量时，计算检验统计值时，变点左边的样本将会非常少，而变点右边纳入到统计量的计算的样本会非常多。而通过图(3.1)-(3.2)得知，当序列处于靠近端点位置时，选取用来计算统计量的样本并非越多越好，而是要依据潜在变点的位置去对应选择纳入到统计量中计算的样本范围。否则的话可能在检验靠近端点附近的变点时，所计算出来的检验统计值偏小而难以拒绝原假设，导致难以识别这一类靠近端点的变点。

因此基于 Zhang，Lavitas[35]和 Dai[8]的研究成果，较好的方法是直接对比可能为潜在变点的j_2前后一部分样本的均值，并且定义对比统计量：

$$D_n(j_1,j_2,j_3)=\frac{(j_2-j_1+1)(j_3-j_2)}{(j_2-j_1+1)^{3/2}}(\overline{X}_{j1,j2}-\overline{X}_{j2+1,j3}) \tag{3.8}$$

因此，$Z_n\left(k^*\right)$可以被重新表示为$D_n(1,k^*,n)$。随后在进行统计量计算时，引入$\lambda_1\in\left(0,\lambda^*\right]$，$\lambda_2\in\left(0,1-\lambda^*\right]$来平衡计算统计量时变点左右的样本量。并且通过之前的讨论，认识到 SN 方法在检验端点处变点的局限性，最终使用检验统计量 LASN，其检验统计量分子部分

$$D_n(k^*-\lfloor n\lambda_1\rfloor,k^*,k^*+\lfloor n\lambda_2\rfloor)\sim\sqrt{n}\lambda_1\lambda_2(\lambda_1+\lambda_2)^{-3/2}\Delta_n$$

其值除了与变点的变化强度有关，同时还与计算统计量时的样本选择情况

有关，为了探究较为直观的关系做出函数 $f(\lambda_1,\lambda_2)=\lambda_1\lambda_2(\lambda_1+\lambda_2)^{-3/2}$ 的等值函数线图。其中 λ_1，λ_2 分别代表变点左右样本选取的比例。

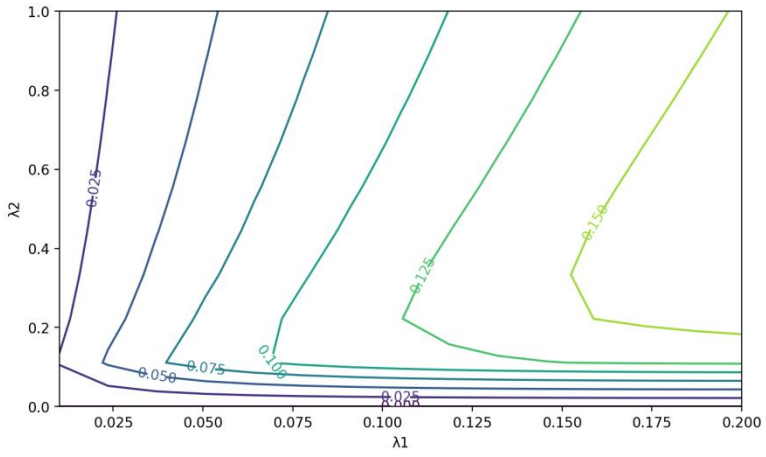

图 3.1：变点 λ 位于序列 0.2 分位点，函数等值线图

通过图 3.1，不难发现，当变点位置位于序列 0.2 分位点时，固定变点左边的样本选取比例 λ_1，$f(\lambda_1,\lambda_2)$ 的值随 λ_2 的增加，其值先增大，后减小。这说明当变点位置靠近左端点时，变点右边用来计算统计量的样本并非越大越好。而较为恰当的选取方式即为后文式子(3.4)计算检验统计量所采取的策略，$\lambda_1=\lambda^*$，$\lambda_1=2\lambda^*$ 是一个不错的样本选取方法。而当变点靠近右端点时，选取方式同理。下面再来研究模拟变点位置位于序列 0.5 分位点时的函数等值线图。

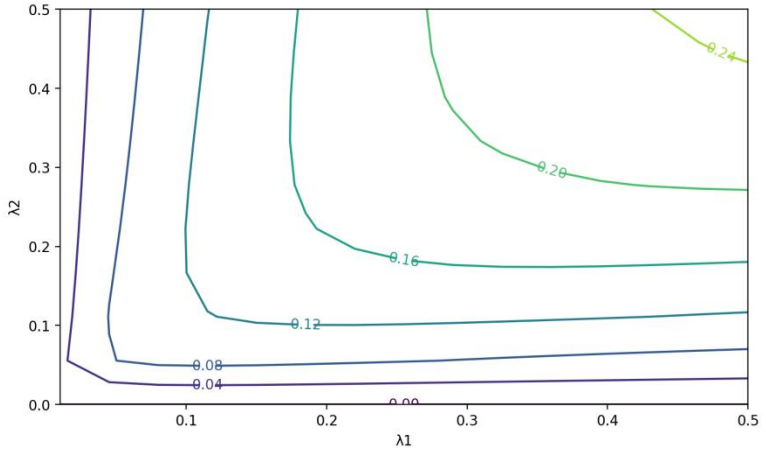

图 3.2：变点 λ 位于序列 0.5 分位点，函数等值线图

通过图 3.2，不难发现当变点位于序列中部时，变点左右样本的选取比例不同于变点位于序列左端时有所限制。等值函数线给直观展示了当变点位于序列中部时，其函数值 $f(\lambda_1,\lambda_2)$ 随着 λ_1，λ_2 的增加而不断地增大，同时也说明了当变点位于序列中部是，在采用 LASN 统计量去检测变点时，潜在变点左右样本地选取比例要尽可能大，这样 LASN 识别变点地能力也就越强。同时，当变点位于序列 0.5 分位点时，LASN 方法的样本选取规则和传统的 SN 方法相同，因此二者在检验变点位于序列中部时的能力也大同小异。

下面再来观察当变点左边的样本选取比例 λ^* 固定为不同的值之后后，$g(\lambda_2)=f(\lambda^*,\lambda_2)$ 的函数图像。

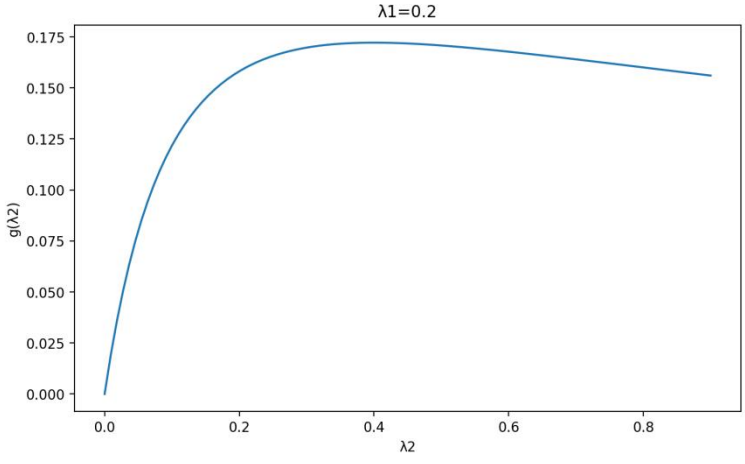

图 3.3：变点 $\lambda^*=0.2$ 时 $g(\lambda_2)$ 的函数图像

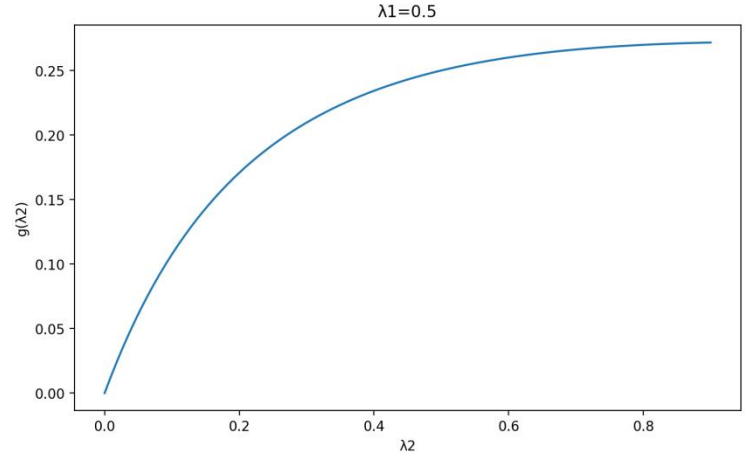

图 3.4：变点 $\lambda^*=0.5$ 时 $g(\lambda_2)$ 的函数图像

图 3.1 和图 3.3 所表达的意思大致相同，还是先来看图 3.3，当变点左侧样本选取比例固定为 $\lambda^*=0.2$ 时，发现函数 $g(\lambda_2)$ 的值并非随 λ_2 的增加而增大，说明了 SN 方法无论计算何处潜在变点统计量时都采用整个序列的做法是一定的局限的。当变点位于靠近左侧端点时，计算统计量时左侧样本选取比例为 0.1 时，右侧样本的最优选择大概为 0.2 时，即 $2\lambda^*$ 时，能够使得 $g(\lambda_2)$ 的值达到最大。

而当变点位置位于序列的中部时，图 3.4 所展示的图像含义和图 3.2 图像说明的含义是类似的。两图都说明当变点位于序列中间时，其左右的样本选择比例越大，函数值或者函数等值线的值都越大，从而在检测变点时都会有更好的效果。

通过上面的讨论，式子 $\lambda_1\lambda_2\left(\lambda_1+\lambda_2\right)^{-3/2}$ 可以通过选择恰当的 λ_1 和 λ_2 来使其达到最大。下面给出当变点位于不同的位置时，统计量计算时样本的选取标准。

$$(1)\quad \lambda^*\in(0,1/3),\lambda_1=\lambda^*,\lambda_2=2\lambda^*$$

$$(2)\quad \lambda^*\in[1/3,2/3],\lambda_1=\lambda^*,\lambda_2=1-\lambda^*$$

$$(3)\quad \lambda^*\in(2/3,1],\lambda_1=2\left(1-\lambda^*\right),\lambda_2=1-\lambda^*$$

上面的关于 λ_1 和 λ_2 的三种取值分别对应了变点可能存在位置的三种情况，第(1)种情况表示了变点位于序列的中间的位置，而(2)和(3)两种情况则代表了变点位于起始位置和末端位置两种情况。可以看出，无论变点位置在初始或者末端位置，选择计算统计量的样本的时候始终遵循如下的规则，需要始终剔除远离可能为变点 k^* 位置的样本，这样 k^* 一侧的样本大小一般是另一个样本的两倍，这保证了一个相对平衡的数据选取。

举个例子，比如 k^* 非常靠近起始端点位置，在构造这个点的检验统计量时会将 k^* 左边的部分全部纳入到计算样本中，而 k^* 右边不会全部选取，不会选取在 $3k^*$ 之后的那些数据。因此，基于位置的自规范化方法进行假设检验构造统计量时，一般不会出现 SN 方法中不管变点 k^* 的位置，永远考虑整个序列，这也有效地提升了检验的功效，虽然有时候这种方法会忽略掉变点前后的变化强度。下面给出更加精细的样本选择的标准：

$\Omega(\varepsilon)=\{(t_1,t_2,t_3):t_1=0,t_3=3\varepsilon,for\ t_2\in[0,\varepsilon);t_1=0,t_3=3t_2,for\ t_2\in[\varepsilon,1/3);$
$t_1=0,t_3=1,for\ t_2\in[1/3,2/3];t_1=3t_2-2,t_3=1,for\ t_2\in(2/3,1-\varepsilon];$
$t_1=1-3\varepsilon,t_3=1,for\ t_2\in[1-\varepsilon,1]\}$

最终，自适应位置的自规范化方法的检验统计量以及其自规范器如下：

$$G_n^{LA}=\max_{(j_1,j_2,j_3)\in\Omega(\varepsilon)}\{D_n^2(j_1,j_2,j_3)/V_n(j_1,j_2,j_3)\} \tag{3.9}$$

其中 $\Omega(\varepsilon)=\{(\lfloor nt_1\rfloor\vee 1,\lfloor nt_2\rfloor\vee 1,\lfloor nt_3\rfloor\vee 1):(t_1,t_2,t_3)\in\Omega(\varepsilon)\}$，

$$V_n(j_1,j_2,j_3)=\sum_{i=j_1}^{j_2}\frac{(i-j_1+1)^2(j_2-i)^2}{(j_3-j_1+1)^2(j_2-j_1+1)}(\overline{X}_{j_1,i}-\overline{X}_{i+1,j_2})^2+$$
$$\sum_{i=j_2+1}^{j_3}\frac{(i-j_3)^2(j_3-i)^2}{(j_3-j_1+1)^2(j_3-j_2)}(\overline{X}_{j_2+1,i}-\overline{X}_{i+1,j_3})^2 \tag{3.10}$$

通过上面的讨论，不难发现最优子样本选择方案 $\Omega(\varepsilon)$ 能够自适应潜在的变化点的位置，更新之后的样本选择方法是显著提高位于端点处的变化点检验功效的关键因素，后面也会通过模拟实验来进一步说明。

3.3 基于位置加权的单变点检测方法

在上面的讨论介绍了根据潜在变点的位置，自适应选择纳入到统计量计算的样本。而在本小节中，在传统的 SN 方法基础上引入一个权重函数，从而重新构建如下的检验统计量：

$$G_n^W=\max_k w^2(k/n)Z_n^2(k)/V_n(k) \tag{3.11}$$

通过降低端点附近差异的权重，当变化点接近边界时，SN 方法的检验功效可以得到提高。目前来说使用最为广泛的权重函数为 $\omega(t)=[t(1-t)]^{-2\kappa}$，其中 $\kappa\in[0,1/2]$。

同时，为了保证检验统计量极限分布的收敛，还需要引入修剪参数。最终基于传统的 SN 方法的带修剪参数的加权版本的检验统计量的形式如下：

$$G_n^W=\max_{k=\lfloor n\varepsilon\rfloor,\ldots,n-\lfloor n\varepsilon\rfloor}\left(\frac{k(n-k)}{n^2}\right)^{-2\kappa}Z_n^2(k)/V_n(k) \tag{3.12}$$

基于位置加权的 SN 方法操作比较直观，用该检验统计量实施假设检验时也能够得到不错的检验功效，但该方法同样存在一定的问题，由于修剪参数

的存在，当变点恰巧位于未被纳入计算的序列内的时候，往往会导致检验的经验功效较差。对于该结论，会在后续的文章中会通过模拟实验给出相应的说明。

因此，最后给出基于位置加权方法的变点的搜寻方法即：

$$(j_1^*, j_2^*, j_3^*) = \underset{(j_1^*, j_2^*, j_3^*) \in \Omega_n(\varepsilon)}{\arg\max} \{D_n^2(j_1, j_2, j_3) / V_n(j_1, j_2, j_3)\} \qquad \textbf{(3.13)}$$

3.4 基于 LASN 方法的多变点检测方法

首先，不失一般性，设置一段存在多个变点的序列，序列由如下形式生成：

$$X_t = \mu 1_{(\lfloor n\lambda_1^* \rfloor < t < \lfloor n\lambda_2^* \rfloor)} + \varepsilon_t$$

其中，其中 $\varepsilon_t = \rho\varepsilon_{t-1} + \eta_t$， ρ 代表序列的自相关系数，$\{\eta_t\}$ 为一族独立同标准正态分布的随机变量。μ 为设定变点后的变化强度，n 为模拟序列的长度。

其检测步骤如下：

第一步：取长度为 m 的序列 $\{X_t\}_{t=1}^m$ 作为初始序列，对该段序利用 3.2 小节中，单变点 LASN 方法进行假设检验。

第二步：若第一步中原假设被拒绝，有

$$(j_1^*, j_2^*, j_3^*) = \underset{(j_1^*, j_2^*, j_3^*) \in \Omega_m(\varepsilon)}{\arg\max} \{D_m^2(j_1, j_2, j_3) / V_m(j_1, j_2, j_3)\} \qquad \textbf{(3.14)}$$

其中 j_2^* 即为初始序列中的估计的变点，同时，返回第一步更新初始序列为 $\{X_t\}_{t=j_2^*+1}^{j_2^*+m}$，否则，若第一步未找到变点，则：

第三步：若不拒绝第一步的原假设，同样更新原序列，作序列的增补 $\{X_t\}_{t=1}^{m+\Delta}$，当序列增补完成后，继续利用新增序列重复第一步假设检验操作。

第四步：最终输出序列中所估计的的变点集合 $\{\hat{k}_1^*, \hat{k}_2^*, ..., \hat{k}_m^*\}$

4.变点检测方法检验效果模拟

4.1 模型设定

在统计学的假设检验理论中，很关键的一点是要设置检验统计量和临界区域，而对于假设检验的效果可以从以下两个方面来入手：

犯第一类错误概率：即拒真错误，指的是在原假设为真时，计算出来的检验统计值却落在拒绝域中，犯第一类错误概率往往被称为检验水平，常用 **size** 来表示。

犯第二类错误概率：即纳伪错误，指的是在原假设为假时，计算出来的检验统计值却落在非拒绝域中，不犯第二类错误概率往往被称为检验功效，常用 **power** 来表示。

对于本章中的模拟实验，对于模拟实验的待检测的序列 $\{X_t\}$ 以如下的方式生成：

$$X_t = \mu 1_{\left(t > \lfloor n\lambda^* \rfloor\right)} + \varepsilon_t$$

其中，其中 $\varepsilon_t = \rho\varepsilon_{t-1} + \eta_t$， ρ 代表序列的自相关系数，而 $\{\eta_t\}$ 为一族独立同标准正态分布的随机变量序列， μ 代表序列的变化幅度。

下面给出五种类型的检测序列，分别用 M1，M2，M3，M4，M5 来表示。

M1： $\varepsilon_t = -0.7\varepsilon_{t-1} + \eta_t$

M2： $\varepsilon_t = -0.3\varepsilon_{t-1} + \eta_t$

M3： $\varepsilon_t = \eta_t$

M4： $\varepsilon_t = 0.3\varepsilon_{t-1} + \eta_t$

M5： $\varepsilon_t = 0.7\varepsilon_{t-1} + \eta_t$

对于每一种模型，在做模拟的时候，每种设定下重复模拟 500 次。

4.2 似然比检验统计量的模拟实验

由于在似然比检验统计量中，事先未加入自规范化器，在通过蒙特卡罗模拟其临界值的时候，为了方便对比不同的权重函数对变点监测是否有差异，根据式子(2.20)-(2.22)，(2.24)，(2.26)，在计算临界值的时候权重函数取了两种不同的形式：

$$\text{(T1)}\quad w(t)=1$$

$$\text{(T2)}\quad w(t)=(t+1)^2$$

其中权重函数(T1)代表并未对原似然比检验统量做加权处理，(T2)代表对原似然比检验统量做了加权处理。依据式(2.20)-(2.22)对三种未加入自规范化器的统计量做蒙特卡罗模拟。

首先在 2000 个格点上模拟随机游走，模拟的区间选择所需要的[0，2]，每个过程模拟 500 次，取 95%分位数；即在(2.20)-(2.22)，中 T 取 1，m 取值为 1000 得到如下的临界值。

当检验统计量为 $\hat{D}_m(k)$，权重函数为 T1 时，临界值 C_α 为 14.202。当检验统计量为 $\hat{D}_m(k)$，权重函数为 T2 时，临界值 C_α 为 3.883。

当检验统计量为 $\hat{Q}_m(k)$，权重函数为 T1 时，临界值 C_α 为 9.262。当检验统计量为 $\hat{Q}_m(k)$，权重函数为 T2 时，临界值 C_α 为 2.641。

当检验统计量为 $\hat{P}_m(k)$，权重函数为 T1 时，临界值 C_α 为 10.146。当检验统计量为 $\hat{P}_m(k)$，权重函数为 T2 时，临界值 C_α 为 2.906。

当检验统计量为 $\hat{D}^{\mathrm{SN}}(k)$，权重函数为 T1 时，临界值 C_α 为 103.572。当检验统计量为 $\hat{P}_m^{\mathrm{SN}}(k)$，权重函数为 T1 时，临界值 C_α 为 82.899。

有了这几种检验统计量在 5%显著性水平下的临界值之后，之后开始模拟当序列中不存在变点时，假设检验发生误报的情况。

4.2.1 似然比检验统计量的经验接受率模拟

在上面的内容中，本文分别通过各个检验统计量的渐近分布求得了各个

统计量在取不同的权重函数的时候的临界值。该部分会对五种类型的数据利用上述的检验统计量计算其在原假设成立的时候，即数据中不存在变点的时候各个类型的统计量发生误报的可能性。

因为当采用不加入自规范化器的统计量的时候，统计量 $\hat{D}_m(k)$，$\hat{Q}_m(k)$，以及 $\hat{P}_m(k)$ 计算长期方差 Σ 时都会涉及到平滑带宽 b_m 的计算，在模拟检验的经验接受率的时候，的平滑带宽 b_m 的选择如下：

对于数据类型为(M3)而言，的平滑带宽选择为 $b_m = \log_{10}(m)$

对于数据类型为(M2)，(M4)而言，的平滑带宽选择为 $b_m = \log_{10}(m^3)$

对于数据类型为(M1)，(M5)而言，的平滑带宽选择为 $b_m = \log_{10}(m^4)$

对于整个序列长度而言，分别模拟两段原始平稳序列长度 $m=100$ 以及 $m=150$ 的数据，其中序列长度为 n={200，300}；即在(2.20)-(2.22)中，T 取 1，m 取值分别为 100 和 150。随后对每一种统计量取权重函数为 T1 的形式，每种情形下分别模拟 500 次，得到如下的经验接受率：

表 4.1：不同检验统计量在权重函数为 T1 时的经验接受率(1-size)

n	model	统计量				
		D	P	Q	D_SN	P_SN
200	M1	0.872	0.730	0.636	0.994	0.992
	M2	0.914	0.782	0.752	0.990	0.990
	M3	0.935	0.958	0.947	0.986	0.964
	M4	0.922	0.812	0.778	0.990	0.986
	M5	0.858	0.731	0.710	0.984	0.982
300	M1	0.900	0.766	0.684	0.990	0.990
	M2	0.924	0.821	0.782	0.990	0.990
	M3	0.954	0.952	0.951	0.978	0.970
	M4	0.932	0.836	0.806	0.989	0.984
	M5	0.882	0.762	0.752	0.986	0.982

经验接受率另外一个角度其实反映的就是假设检验不犯第一类错误的一个情况，即原假设为真的时候，不拒绝原假设的一个概率。透过模拟的经验接受率的表，可以从以下几个方面来分析。

首先就是从原平稳序列的长度以及待检测序列的长度 m 来看，无论是未加入自规范化器的统计量 $\hat{D}_m(k)$ 还是 $\hat{Q}_m(k)$，抑或是 $\hat{P}_m(k)$，还是加入了自规范化器的 $\hat{D}^{\mathrm{SN}}(k)$ 或者 $\hat{P}_m^{\mathrm{SN}}(k)$。当初始平稳序列较小时比如 m 取值为 100 的时

候，其经验接受率相比于当初始平稳序列较大时比如 m 取值为 150 的时候是偏小的。因此恰当增加最初的原始序列长度是有助于减小的假设检验犯第一类错误的概率的。

其次再从数据的结构上来看各个统计量的效果，由于知道模型 M3 是无结构相关性的数据，而模型 M2，M4 和模型 M1，M5 是 AR(1)类型的数据，两者区别是后面两个模型的自相关系数取得较大为-0.7 和 0.7，模型 M2，M4 自相关系数取得较小为-0.3 和 0.3。模型 M3 则是正态分布类型的数据。从数据的相关性结构上来看，无结构相关性的数据 M3 做假设检验得到的犯第一类错误的概率是较为接近取临界值时的显著性水平的，而 AR(1)类型的数据做假设检验时的经验接受率往往是低于无结构相关性的数据 M3，因此，上述检验统计量检验 AR(1)类型的数据时，发生误报率的概率直观上来看是要高于无结构相关性的数据的。且随着 AR(1)类型的数据自相关系数的增大，检验犯第一类错误的概率是在增大的。

最后，再来看统计量加入自规范化器之后的效果，首先很直观的就是在加入了自规范化器之后相比于用长期方差，犯第一类错误的概率是明显减小的。通过学者的研究这主要是由于在使用长期方差的时候平滑带宽 b_m 的选择很大程度上是人为自己所选择的，有时候其很难反映出真实的数据结构从而导致误报率相对于使用自规范化器而言相对较高。此外对于检测 AR(1)类型的数据，自规范化类型的统计量的经验接受率直观上是高于直接使用似然比统计量的，因此能够很好地减小假设检验犯第一类错误的概率。

4.2.2 似然比检验统计量的经验检验功效

之后为了模拟似然比检验统计量变点检测的经验检验功效，需要对变点发生时刻所在的位置做一些假定，变点出现的位置设定如下：

$$X_t^{\mu}=\begin{cases} X_t & \text{if } t < m+\lfloor 0.9m \rfloor \\ X_t+\mu & \text{if } t \geq m+\lfloor 0.9m \rfloor \end{cases}$$

同样的，为了探索不同长度的原始序列对检验的经验功效是否有影响，在模拟检验的经验功效时同样采取两种原始序列假定，$m=100$ 以及 $m=150$，序列总长度则为 200 和 300。因此变点的位置是位于 190 和 285 这两个位置。

对于变化的强度μ，在模拟实验中，为了模拟序列从小到大的一个变化情况，的变化的强度μ取值分别为0.1，0.5，1.0，2.0，3.0其他的数据类型还是同前面模拟其经验接受率一致，其具体形式如4.1节中的模型假定：

每个数据类型的变点分别为190和285，每个数据类型，每个变化强度μ以及每一个权重函数T，每个情形下分别模拟500次，得到如下的经验检验功效。

表4.2：不同检验统计量在权重函数为T1，n=200时的经验检验功效(power)

统计量	模型	n = 200，T=T1				
		0.1	0.5	1	2	3
D	M1	0.150	0.280	0.704	1.000	1.000
	M2	0.088	0.136	0.392	0.968	1.000
	M3	0.052	0.076	0.176	0.766	1.000
	M4	0.080	0.100	0.138	0.440	0.834
	M5	0.128	0.126	0.130	0.194	0.294
P	M1	0.316	0.494	0.862	1.000	1.000
	M2	0.214	0.276	0.626	1.000	1.000
	M3	0.046	0.068	0.106	0.304	0.746
	M4	0.182	0.186	0.276	0.674	0.954
	M5	0.262	0.260	0.258	0.338	0.494
Q	M1	0.364	0.544	0.890	1.000	1.000
	M2	0.248	0.356	0.686	1.000	1.000
	M3	0.046	0.062	0.100	0.256	0.508
	M4	0.224	0.228	0.356	0.722	0.970
	M5	0.292	0.296	0.294	0.388	0.534

根据上面各统计量在不同变化幅度下的经验功效表，可以从以下几个角度来分析：

最直观的一点就是对于统计量D，统计量P，统计量Q，无论数据类型是存在结构相关性还是不存在结构相关性，每一种统计量在每一种结构的数据之下都具有这样的一种性质，即随着变化强度μ的不断增大，各种统计量对于变点的识别能力都是提升的。具体来看无论是统计量D，还是统计量P，或者统计量Q，当变化强度μ小于0.5的时候，无论哪一种结构的数据还是哪一种方法，通过对应的检验统计量进行检验的检验功效都小于0.5的。此时，以上三种方法识别变点的能力是比较弱的。