BIOMETRICS ??, 1–18 DOI: 10.1111/j.1541-0420.2005.00454.x
??? ????

多元变点检测的非渐近推断

伊恩·巴内特

美国宾夕法尼亚州费城，宾夕法尼亚大学生物统计学系

电子邮箱：ibarnett@pennmedicine.upenn.edu

arXiv:2304.04644v1 [stat.ME] 10 Apr 2023

SUMMARY: 变点检测中的传统推断方法通常依赖大量观测数据，在非渐近情形下可能并不准确。随着移动健康和数字表型研究的兴起，患者通过智能手机或其他数字设备接受监测，因此需要在非渐近情形下进行变点检测，以识别诸如复发或自杀等不良事件发生前仅数日出现的行为变化。此外，大规模数字表型队列的监测和在线分析需要分析和计算效率高的推断方法。我们将似然比检验渐近尾概率的结果扩展到多元变点检测情形，并通过模拟说明当观测数据数量不大时，该结果并不准确。我们提出一种用于似然比检验推断的非渐近方法，并将其估计的 P 值效率与通过模拟零分布得到的常用经验 P 值进行比较。通过模拟，以及检测一名精神分裂症患者复发前一周的行为变点，我们展示了该方法相较于其他方法的准确性和检验效能。KEY WORDS: 变点检测；数字表型分析；移动健康；非渐近推断；序贯分析。

本文已提交，正在考虑是否发表于生物统计学

1. Introduction

本工作的主要目标是在不依赖渐近理论或在零分布下进行计算密集型模拟的情况下，准确估计多元变点检测似然比检验统计量的尾部概率。这一问题出现的背景是，越来越多的研究依赖于数字表型分析，Torous 等人（2016）将其定义为“利用来自智能手机和其他个人数字设备的数据，对个体层面的人类表型在现场进行逐时刻量化”。数字表型分析正越来越多地用于研究人类行为，主要用于监测患有精神疾病的脆弱人群 (OnnelaandRauch, 2016;Torous et al.,2017; Insel,2017)。典型的数字表型分析研究会让患者在自己的个人智能手机上安装一个应用程序，该程序会在日常使用期间于后台收集传感器数据。这些数据可用于计算与患者位置、移动性、活动、社交性和睡眠相关的行为特征，并在较长时间内以高粒度进行分析。将单个患者的这些行为特征视为多元时间序列后，该时间序列中的变点可能对应于患者行为的改变，或症状加剧并最终导致复发或自杀等不良事件。由于变点检测所需的变点后观测数据点很少，推断的渐近方法并不适用

尽管已有大量研究致力于开发强大的多元变点检测方法，但许多研究仅关注识别最可能的变点位置，却未能解决相应的统计推断问题，且不使用对零分布进行蒙特卡罗模拟的方法。例如，基于累积和控制图（CUSUM）的方法（Cho et al., 2016; Cho and Fryzlewicz,2015）、成对似然方法（Ma and Yau, 2016）、基于似然的代价函数的推广（Killick et al.,2012），以及允许存在变点的方法

在特征的协方差方面（LavielleandTeyssiere, 2006；Kirch et al., 2015），人们已经开发出非参数方法（Preuss et al., 2015)，例如基于样本散度的方法（Matteson and James, 2014）、一种允许存在缺失数据的降维方法（Xie et al., 2013），以及一种能够放宽“变点在多元时间序列中同时发生”这一假设的方法（Bardwell et al., 2018）。尽管这些方法旨在针对特定备择假设有效检测变点，但在每种情况下，非渐近统计推断问题几乎都被忽略，通常完全依赖对零分布进行蒙特卡罗模拟来开展假设检验。虽然在这些方法所考虑的应用场景中，模拟零分布或许可以接受，但在本文中，我们量化了这种推断在显著性水平较低的情况下可能带来的高昂计算成本。由于需要针对长期随访期间反复开展在线分析进行多重比较校正，小显著性水平的问题存在于基于智能手机的数字表型研究中。为解决这一问题，本文针对均值的多元变点检测中最常用且最基础的方法之一——似然比检验，开发了一种非渐近推断方法。

似然比变点检测检验统计量的渐近尾概率和 P 值估计已得到研究，Siegmund（2013）对各类变点问题进行了全面的渐近分析。由于随机过程弱收敛，随访时间足够长时，数字表型可用布朗运动近似，据此可获得尾概率（Billingsley，2008）。对于单变量过程，研究者已针对检测一个变点（Woodroofe，1982）以及两个变点（即一个变点区间）（Siegmund，1992；Hogan 和 Siegmund，1986）开发了这些尾概率的渐近近似。James 等人（1987）探讨了将方法扩展至多元过程，用于检测单个变点；

Zhang 等人（2010）；Siegmund 等人（2011）则研究了变点区间，在这些研究中，变点检测用于检测拷贝数变异，因此采用渐近理论是合适的。数字表型研究中可获得的观测数据要少得多。例如，如果每天收集行为特征，那么只有在变点发生后几天内检测到变点，利用变点检测来触发干预才有意义。统计过程控制所用的 CUSUM 程序（Page（1954））考虑了非渐近情形下的变点检测问题，但针对多元变点检测问题的似然比检验，其非渐近分析仍鲜有研究。Kiefer（1959）利用贝塞尔函数的无穷级数推导出了多元情形下的精确尾概率，但这种方法会在整个过程中搜索变点，而我们只关注对患者近期病史中的变点进行有针对性的检验。

本文其余部分的组织如下。我们首先在第2 节介绍均值变点模型。在第3节中，我们推导多元变点检测的渐近概率，并将其与我们提出的非渐近推断方法进行比较，展示其在非渐近情形下的较差准确性。在第 4 节中，我们从理论和模拟两方面，将我们提出的 p 值估计的效率与通过对零分布进行蒙特卡罗模拟获得的结果进行比较。证明留在附录中。在第 5节中，我们使用检验效能模拟来比较竞争方法；在第6 节中，我们将所提出的方法应用于检测一名精神分裂症患者复发前一周数字表型中的变点，从而展示多元变点检测在数字表型研究中的潜力。

2. 模型

我们考虑服从以下分布的观测数据 Y_{ij}：

$$Y_{ij} = \mu_i + I_{\{j>k\}}\delta_i + \epsilon_{ij} \tag{1}$$

对于特征 $i = 1$、.....、 q 以及按时间顺序排列的观测值 $j = 1$、.....、 n，其中 $I_{\{x\}}$ 是 x 的指示函数， ϵ_{ij} 是相互独立且同分布的标准正态误差。Zhang 等人（2010）考虑了一个类似的变点检测模型，其中变点区间（k_0、 $k_1]$ ）具有某个最小宽度，使得 $1 \leqslant k_0 < k_1 \leqslant n$ 指定均值发生 δ_i 偏移的特征 i。模型 1 表示其中 $k_1 = n$ 的特殊情形。之所以对 k_1 作此限制，是因为在使用变点检测来提示对脆弱人群进行干预时，识别近期的行为变化才是重要的。有效的干预需要对行为变化作出及时响应，而患者过去发生的行为变化区间（$k_1 < n$）意味着患者的行为最终已恢复正常状态，因此不再需要干预。

对于特征 i，在时间 k 检验变点对应于 H_0： $\delta_i = 0$ 以及双侧备择 H_1： $\delta_i \neq 0$。令 $\bar{Y}_i = \sum_{j=1}^{n} Y_{ij}/n$，Olshen 等人（2004）和Zhang等人（2010）针对单个特征 i 在时间 k 发生变点所提出的基于似然比的检验统计量为：

$$U_{ik} = \frac{(\sum_{j=k+1}^{n} Y_{ij}) - \bar{Y}_i(n-k)}{\sqrt{k(n-k)/n}} \tag{2}$$

该检验统计量的性质很直观：如果 Y_{ij} 的均值在时间 k 发生变点，那么变点之后的观测数据总和应与所有数据的观测平均值 Y_i相比有显著差异¯。为了检验所有 p 个特征在时间 k是否同时发生变点，我们考虑如下检验统计量：

$$Z_k = \sum_{i=1}^{q} U_{ik}^2 \tag{3}$$

对于每个 k、 $Z_k \sim \chi_q^2$ 在 H_0 下的情况，这对应于 $\delta_1 = \cdots = \delta_q = 0$ 时的情形。由于

变点位置未知时，我们通过考虑以下检验统计量，在区间 $[m_0, m_1]$ 内搜索最可能的变点位置：

$$Q = \max_{n-m_1 \leqslant k \leqslant n-m_0} Z_k$$

当 Q取较大值时，我们拒绝 H_0 。

3. P 值估计方法

对于较大的 n，Siegmund（2013）在单变量情形（$q = 1$）下推导出了P 值pr($Q \geqslant b^2$)的近似，该近似可扩展到任意 q 特征，如下所示：

THEOREM 1: 假设 $b \to \infty$、 $m_1 = o\,(n)$ 、 $m_0 \to \infty$ 和 $m_1 \to \infty$ ，使得 $bm_1^{-1/2} = c_1 < c_0 = bm_0^{-1/2}$ 对于常数 c_0、 c_1。那么在 H_0 马会

$$pr(Q \geqslant b^2) \approx 2^{-q/2} \left[\Gamma\left(q/2\right)\right]^{-1} \ln\left(m_1/m_0\right) b^q \exp\left(-b^2/2\right) \tag{4}$$

对于使用每日观测进行智能手机监测的目的而言，我们经常会关注这样的情况：比如，$m_0 = 0$ 和 $m_1 = 6$。这相当于寻找患者在前一周内行为的变点。由于 m_0 和 m_1 被固定为较小的值，即使对一个人进行了长时间的跟踪（$n \to \infty$），渐近近似也不会得到改善。推断必须转而依赖非渐近方法。

作为对这种渐近近似的替代方案，我们详细说明如何在不假设 b、 n、 m_0 和 m_1 较大且不要求对零分布进行计算成本高昂的模拟的情况下，估计 pr($Q \geqslant b^2$)。一种有用的近似是

$$\operatorname{cor}(Z_{j_1}, Z_{j_2}) \approx (n - j_2)/(n - j_1) \quad (1 \leqslant j_1 < j_2 \leqslant n). \tag{5}$$

¯当我们以 Y_i 为条件时，这种相关性是精确的，这意味着对于 $n - m_1 \leqslant j_1 < j_2 \leqslant n - m_0$ ，当 m_1 和 m_0 都接近 0 时，近似成立得很好；对于数字表型研究中所关注的变点位置，情况正是如此。令 $F_q(\cdot)$表示

一个 χ_q^2 随机变量的累积分布函数，我们将 $Z_k^* = \Phi^{-1}\{F_q(Z_k)\}$ 定义为 Z_k 的一个变换，使其在 H_0下服从正态分布。令

$$Q^* = \max_{n-m_1 \leqslant k \leqslant n-m_0} Z_k^*,$$

当 H_0 的值较大时拒绝的检验，与当 H_0 的值较大时拒绝的检验等价，因为 Z_i^*是 Z_i的单调函数。令 $\boldsymbol{Z}^* = (Z_{n-m_1}^*, \ldots\ldots, Z_{n-m_0}^*)$，我们用均值为$\mathbf{0}$ 、协方差为$\boldsymbol{\Sigma}$的多元正态分布来近似 $\boldsymbol{Z}^*$ ，其中 $\boldsymbol{\Sigma}$ 可以通过经验方法近似，或通过对 Z_i^*进行泰勒级数展开解析地求得。请注意，尽管每个 Z_i^*的边际分布均为正态分布，但由于其依赖结构以及用于得到 Z_i^* 的非线性变换，将多元正态分布用于它们的联合分布时只能是一种近似。尽管如此，接下来我们将证明，多元正态近似相当准确。

$\mathrm{cor}(Z_{j_1}^*, Z_{j_2}^*)$的一阶泰勒级数近似为$(n-j_1)/(n-j_2)$，其中 $j_1 < j_2$。 ˆ 一旦估计出 $\boldsymbol{\Sigma}$，即可对 $MVN(\mathbf{0}, \boldsymbol{\Sigma})$分布进行数值积分，从而得到：

$$\mathrm{pr}(Q \geqslant b^2) = 1 - \mathrm{pr}\left[Q^* < \Phi^{-1}\{F_q(b^2)\}\right].$$

正如第 4 节所示，经验估计 $\boldsymbol{\Sigma}$ 比直接经验估计 $\mathrm{pr}(Q \geqslant b^2)$ 更容易，因为对 Q 的推断对 $\boldsymbol{\Sigma}$ 的估计相当稳健，因此，与直接对零分布进行蒙特卡罗模拟所需的模拟次数相比，只需更少的模拟次数即可获得准确的 p值。

[Table 1 about here.]

为展示在与大多数数字表型研究相似的情境中渐近 p值的不准确性，我们根据模型1进行模拟，在 H_0 with $\delta_i = 0$ andwith$n = 30, 100, 365,$ and 1000天的数据收集量下进行，这些数值分别表示收集的数据天数，以及 $q = 5$,10 和 50，表示每天收集的特征数量。我们比较了三种获取 p值的方法：根据方程 4 得到的渐近 p值、基于

Q^* 使用多元正态分布累积分布函数的数值积分，其中$\mathbf{\Sigma}$ 通过经验方法估计；另一种方法则采用相同的做法，但使用公式 5 给出的其一阶近似来估计 $\mathbf{\Sigma}$ 。表 1 展示了这些方法在各种设定下、显著性水平为 0.05 时的第一类错误率。 $m_0 = 0$ 和 $m_1 = 6$设定对应于数字表型研究，旨在识别最近一周内的变点；而 $m_0 = \sqrt{n}/2$ 和 $m_1 = \sqrt{n}$ 设定则反映了传统的渐近变点检测方法。在 $m_0 = \sqrt{n}/2$ 和 $m_1 = \sqrt{n}$ 设定下，渐近近似有显著改善，尤其是在 n 较大且 q较小时。在 $m_0 = 0$ 和 $m_1 = 6$ 设定下，即使 n 非常大，渐近 P 值仍会不准确，因为所有渐近变点检测方法都要求 m_0 和 m_1 同时增大。相比之下，依赖于估计 $\mathbf{\Sigma}$ 的两种方法，在 $m_1 =$−6 和 $m_0 = 0$ 时比在 $m_1 = \sqrt{n}$ 和 $m_0 = \sqrt{n}/2$时更加准确，因为 $\mathbf{\Sigma}$ 的维数为 $m^* \times m^*$，其中 $m^* = m_1 - m_0 + 1$，因此 m^* 较小时更容易进行估计。总体而言，经验估计 $\mathbf{\Sigma}$ 是最准确的方法；由于其联合分布采用了多元正态近似，该方法的表现略偏保守。在 $m_1 = 6$ 和 $m_0 = 0$ 情形下， $\mathbf{\Sigma}$ 的一阶近似表现也可以接受，尽管更加保守。鉴于其准确性较高，接下来的检验效能模拟和数据分析部分将采用经验估计 $\mathbf{\Sigma}$ 的方法，对似然比检验进行推断。

4. 基于经验的 p值的准确性

[图 1 置于此处。]

对于给定的临界值 a 检验统计量 Q^*，令相应的 P 值为 $p =$pr$(Q^* \geqslant a) = 1 - f(\mathbf{\Sigma})$, 其中

$$f(\mathbf{\Sigma}) = \frac{1}{\sqrt{|\mathbf{\Sigma}|(2\pi)^{m^*}}} \int_{-\infty}^{a} \cdots \int_{-\infty}^{a} \exp\left(-\frac{1}{2}\boldsymbol{y}^T \mathbf{\Sigma}^{-1} \boldsymbol{y}\right) d\boldsymbol{y}.$$

由于未观测到 $\mathbf{\Sigma}$ ，为了计算该积分，我们必须改为估计 $\mathbf{\Sigma}$ 。假设我们为 $b = 1$、.. . 、B 各生成 $\boldsymbol{Z}^{*(b)}$ ，每个都在 H_0 下。令 $\hat{\mathbf{\Sigma}}$ 为$\boldsymbol{Z}^{*(b)}$的经验相关性，在第 3 节中我们提出使用

$$\hat{p} = 1 - f\left(\hat{\mathbf{\Sigma}}\right)$$

以估计 P 值。这是直接使用经验方法估计 P 值的另一种替代方案

$$\tilde{p} = \sum_{b=1}^{B} I_{\{Q^{*(b)} \geqslant a\}}/B.$$

$\tilde{p}$ 的方差为 $p(1-p)/B$ ，因为它是一个二项比例。作为比较，我们求$\hat{p}$的方差:

THEOREM 2: 鉴于 $m^* = m_1 - m_0 + 1$ 是 $\boldsymbol{Z}^*$ 的长度，并令 $\mathbf{\Sigma}$ 为基于 H_0 下 $\boldsymbol{Z}^*$ 的 B 个模拟实例所得的 $\mathbf{\Sigma}$ 的经验估计量。令$\hat{p} = 1 - f\left(\hat{\mathbf{\Sigma}}\right)$，则:

$$var(\hat{p}) = c_a \cdot \frac{p^2}{B} + o_B\left(B^{-1}\right) \tag{6}$$

其中

$$\begin{aligned} c_a = & \sum_{1 \leqslant i < j \leqslant m^*} \left\{\left(\mathbf{\Sigma}^{-1}\right)_{ij} - h_{ij}(a)\right\}^2 \left(1 - \Sigma_{ij}^2\right)^2 \\ & + 2 \sum_{1 \leqslant i < j < k \leqslant m^*} \left\{\left(\mathbf{\Sigma}^{-1}\right)_{ij} - h_{ij}(a)\right\} \left\{\left(\mathbf{\Sigma}^{-1}\right)_{jk} - h_{jk}(a)\right\} \left(\Sigma_{ik} + \Sigma_{ij}\Sigma_{jk}\right) \end{aligned}$$

and

$$\begin{aligned} h_{ij}(a) &= \tfrac{1}{1 - f(\mathbf{\Sigma})} \int\limits_{\boldsymbol{y}: \max_j y_j > a} \frac{1}{2} \boldsymbol{y}^T \mathbf{\Sigma}^{-1} \frac{\delta \mathbf{\Sigma}}{\delta \Sigma_{ij}} \mathbf{\Sigma}^{-1} \boldsymbol{y} \frac{\exp\left(-\frac{1}{2} \boldsymbol{y}^T \mathbf{\Sigma}^{-1} \boldsymbol{y}\right)}{|\mathbf{\Sigma}|^{1/2} (2\pi)^{m^*/2}} d\boldsymbol{y} \\ &= E\left(\tfrac{1}{2} \boldsymbol{y}^T \mathbf{\Sigma}^{-1} \tfrac{\delta \mathbf{\Sigma}}{\delta \Sigma_{ij}} \mathbf{\Sigma}^{-1} \boldsymbol{y} \middle| \max_j y_j > a\right). \end{aligned}$$

虽然公式 6 中的 c_a 项随 p 递增（当 $p \to 0$时），但其增长速率至多为$(\log p)^2$ ，因此 var($\hat{p}$)主要由 p^2 项决定。定理2的以下推论通过直接比较$\hat{p}$ 与$_{\sim}p$的方差说明了这一点。

COROLLARY 1: 存在 c^* ，使得对于所有 $p < 0.5$ ，我们都有以下关系:

$$\frac{var(\hat{p})}{var(\tilde{p})} \leqslant c^* \cdot \frac{p(\log p)^2}{(1-p)} + o_B(1) \tag{7}$$

方程 7 的含义是，对于相当小的 P 值，$\hat{}p$ 是远优于 $\tilde{p}$ 的估计量。然而，在 c_a 较大且 p 不小的情况下，$\tilde{p}$ 可能比 $\hat{}p$ 更有效地估计 p 值。由于 $\mathbf{\Sigma}$ 是 $m^* \times m^*$，当 m^* 较大时，经验估计量 $\hat{}\,\mathbf{\Sigma}$ 的噪声较大，这会导致更大的 c_a，以及 $\hat{}p$ 更高的变异性。

一个好的 P 值估计量应有很高的可能性接近真实的 p。作为比考察变异性更易于解释的替代方法，我们通过模拟研究比较 pr($|\hat{}p - p| < \delta$) 与 pr($|\tilde{p} - p| < \delta$)：令 $p = 0.05$，0。001；对于给定的 B，使用 1000 次迭代来确定 $\hat{}p$ 和 $\tilde{p}$ 同时处于 p 的 5% 范围内（即 $\delta = 0.05p$）的比例。模拟结果见图 1。罕见事件二项比例的离散特性解释了 $\tilde{p}$ 的不稳定性。显然，$B = p^{-1}$ 至少需要经过若干次迭代，$\tilde{p}$ 才能得到正确的 P 值，即便如此，成功的可能性仍然很小。只有当 B 比 p^{-1} 大许多个数量级时，$\tilde{p}$ 才会成为 p 的实用估计量。相比之下，即使 B 非常小，$\hat{}p$ 仍是 p 的可靠估计量。将 m^* 从 $m_1 = 3$ 增至 $m_1 = 6$ 会增加 c_a，从而降低 $\hat{}p$ 的准确性，但 pr($|\hat{}p - p| < \delta$) 的下降相对较小，且仅在 B 较小时才会产生影响。此外，$\hat{}p$ 的表现对 p 的大小具有稳健性，因此是中等偏小至极小 p 的理想估计量。

5. 数值功效比较

我们还考虑了几种用于多元变点检测问题、旨在检验模型1 的似然比检验替代方法。Hotelling 的多元 T^2 检验使用统计量 $\sum_{i=1}^{q} \hat{}\epsilon_{ij}^2$，以检验第 j天是否存在变点，其中 ϵ_{ij} 是在 H_0 下根据模型1 估计得到的（Hotelling，1947）。Crosier（1988）提出的累积控制和图程序的多元扩展设定目标值 $a > 0$，并令 $C_j =$

$\{\sum_{i=1}^{q}(s_{ij} + \hat{\epsilon}_{ij} - a)^2\}^{1/2}$ 其中，当 $s_{ij} = 0\,(i = 1, \ldots, q)$时，$C_j \leqslant \kappa$ 对于某个阈值$\kappa > 0$，否则为 $s_{ij} = (s_{i(j-1)} + \hat{\epsilon}_{ij} - a)(1 - \kappa/C_j)$。该过程初始化为$s_{i0} = 0\,(i = 1, \ldots, q)$，并且当 $\max_j(\sum_{i=1}^{q} s_{ij}^2)$取较大值时拒绝 H_0。通过对每个$\hat{\epsilon}_{ij}$ 的符号取反并重复该过程，可以将该检验扩展到双侧假设。该检验与Hotelling的 T^2有相似之处，但不同的是，它旨在累积多个连续时间点上的均值变化。

[图 2 放置于此处。]

为了比较这些方法在旨在模拟数字表型研究的非渐近情形下的表现，我们考虑观测 $n = 100$ 天，变点位置 k 出现在随访结束（第 n 天）前一周内。所考虑的特征数量为 $q = 10$ 和 $q = 20$，并根据模型 1 模拟数据，其中 $\delta_i = 0.5$。多元累积控制和图法采用了两个不同的阈值，分别为 $\kappa = 2$ 和 $\kappa = 3$。针对每种参数设置和每个变点位置，基于 1000 次迭代估计检验效能，结果见图 2。

始终是似然比检验的功效最高，其次是多元累积控制和图法，再次是Hotelling 的 T^2。随着变点进一步向过去移动，每种方法的功效都会提高，这直观上对应于变点之后观测数量的增加。随着特征数量增加，多元累积和控制图法得益于使用更小的阈值 κ。将特征数量从 $q = 10$ 增加到 $q = 20$，持续提高了所有方法的功效，并展示了在非渐近变点问题中采用多元且多样化特征集的必要性。在变点之后观测数量如此之少的情况下，检测变点的唯一可行方式是该变点同时出现在各种特征中。在这些情形下，单变量变点检测的检验效能将严重不足。

6. 用变点检测预测复发

[图 3 见此处。]

我们将一名精神分裂症患者的每日行为特征视为一项数字表型分析试点研究的一部分；该研究从马萨诸塞州波士顿的一家州立心理健康诊所招募受试者。在招募期间，该患者在其个人智能手机上安装了 Beiwe应用程序；在后续随访期间，通过后台使用收集了 GPS、加速度计、通话记录及其他数据流（Torous等，2016）。我们考虑一组代表患者行为每日汇总的七项特征，这些行为与移动性、活动和社交性相关：在家中度过的小时数、行进距离（千米）、距家的最大距离（千米）、一天中未移动时间所占的比例、通话出度、通话互惠性以及昼夜节律惯例。昼夜节律惯例根据每日GPS 数据进行量化，详见 Canzian 和 Musolesi（2015）。

经过六周随访后，患者复发并住院。本文回顾性地对这些特征应用多元变点检测方法，以确定是否可以根据患者住院前行为的变化预见复发。仅考虑复发前观测到的 39 天数据，每种变点检测方法都检测了复发前一周内这些特征是否出现变点。Hotelling 的 T^2 以及同时采用 $\kappa = 2$ 和 $\kappa = 3$ 的多元累积和控制图法均未能检测到这些特征中显著的变点，但似然比检验检测到复发前两天行为发生变化，且具有边缘显著性（$p = 0.041$）。

图 3展示了复发前行为特征及其变化的可视化表示。在复发前两天，患者拨打电话的联系人数量明显增加，回拨率上升，日常行为出现明显偏离，待在家外的时间增加，并且前往距离住所比平时更远地点的次数增多。

尽管单独来看，每项特征偏离均值的程度未必值得注意，但它们同时发生变化，反映出一种更广泛的模式，使得基于这些信号的组合能够实现检测。尽管这一结果令人鼓舞，但在实际应用中，由于需要进行多重比较校正，仅有边际显著性是不够的。最可靠地提升统计功效的方法，是扩充数字表型分析的特征集，以更好地量化行为。

7. 讨论

对于非渐近单变量变点检测而言，需要非常强的信号，才能从相对少量的观测中识别出变化。因此，多元变点检测能够汇总不同特征的信息，即使每个特征的边际变化都较小，也能检测到变点。第 6 节中介绍的七种行为特征并非面面俱到，也无法充分体现从智能手机数字表型数据中提取的有用信息。随着数字表型研究逐渐超越可行性研究和试点研究阶段，我们需要开发标准化方法和软件，以清理数据，并将其量化为更全面且有意义的一组行为特征，涵盖移动性、睡眠、活动和社交性。随着 q 增加，可能有必要进行降维，以减少特征集中的冗余，同时提高检验效能。

虽然第 6节中 प्रस्तुत的数据示例是一项回顾性分析，但最终有必要将变点检测方法应用于在线方式，以便在不良事件发生之前检测到行为变化，从而及时采取干预。数字表型研究中许多受到关注的心理健康结局都很罕见，例如自杀未遂或复发，因此需要开展新的大规模数字表型研究，以探究多元变点检测方法的敏感度和特异度。这将有助于确定实践中应采用的适当显著性水平，从而同时

尽量减少假阳性的出现频率，同时保持足够的灵敏度，以检测细微的行为变化。事实上，尽管多元变点检测是至关重要的第一步，但对于检测到变点后何为理想的干预或应采取何种行动，目前尚未达成共识。

Supplementary Materials

第 6 节中使用的精神分裂症数字表型，以及用于复现所有结果的代码，可在 Wiley Online Library 上的《生物统计学》网站与本文一同获取。

References

Bardwell, L., Fearnhead, P., Eckley, I. A., Smith, S., 和 Spott, M. (2018). 面板数据中最新的变点检测。计量学 第 1–28 页。Billingsley, P. (2008). 概率与测度。John Wiley & Sons。

Canzian, L. 和 Musolesi, M. (2015). 抑郁症轨迹：通过分析智能手机移动性轨迹，对抑郁状态进行无干扰监测。载于*2015* 年 *ACM* 普适与泛在计算国际联合会议论文集，第 1293–1304 页。ACM。

Cho, H. 等 (2016). 基于双重累积和统计量的面板数据变点检测。电子统计学期刊 **10,** 2000–2038。

Cho,H.和 Fryzlewicz, P.（2015）。通过稀疏化二分分割进行高维时间序列的多变点检测。英国皇家统计学会杂志：*B*辑（统计方法学）**77,** 475–507。

Crosier, R. B.（1988）。累积和质量控制方案的多变量推广。*Technometrics* **30,** 291–303。

Hogan, M. 和 Siegmund, D.（1986）。某些随机场最大值的大偏差。应用数学进展 **7**,2–22。

Hotelling, H.（1947）。多元质量控制。统计分析技术 。Hotelling, H.（1953）。关于相关系数及其变换的新见解。英国皇家统计学会学报：*B*辑（方法论） **15**,193–232。Insel, T. R.（2017）。数字表型分析：用于行为科学新领域的技术。*Jama***318**, 1215–1216。James, B., James, K. L. 和 Siegmund, D.（1987）。变点检验。*Biometrika* **74**,71–83。Kiefer, J.（1959）。Kolmogorov-Smirnov 检验和 Cramér–von Mises 检验的 K 样本类比。数学统计年刊 第 420–447 页。Killick, R., Fearnhead, P. 和 Eckley, I. A.（2012）。以线性计算成本实现变点的最优检测。美国统计学会杂志 **107**,1590–1598。Kirch, C., Muhsal, B. 和 Ombao, H.（2015）。多元时间序列变化的检测及其在脑电图数据中的应用。美国统计学会杂志 **110**,1197–1216。Lavielle, M. 和 Teyssiere, G.（2006）。多元时间序列中多个变点的检测。立陶宛数学杂志 **46**, 287–306。Ma, T. F. 和 Yau, C. Y.（2016）。基于成对似然的多元时间序列模型变点检测方法。*Biometrika* **103**, 409–421。Matteson, D. S. 和 James, N. A.（2014）。多元数据多重变点分析的非参数方法。美国统计学会杂志**109**,334–345。Nardi, Y., Siegmund, D. O., Yakir, B. 等（2008）。近似高斯随机场最大值的分布。统计年刊 **36**, 1375–1403。Olshen, A. B., Venkatraman, E., Lucito, R. 和 Wigler, M.（2004）。环形二元

用于分析基于芯片的 DNA 拷贝数数据的分割。生物统计学**5,** 557–572。

Onnela, J.-P. 和 Rauch, S. L.（2016）。利用基于智能手机的数字表型分析，提升行为与心理健康研究。《神经精神药理学：美国神经精神药理学学会官方出版物》**41,** 1691–1696。

Page, E. S.（1954）。连续检验方案。《*Biometrika*》**41,** 100–115。

Preuss, P., Puchstein, R. 和 Dette, H.（2015）。多变量时间序列中多个结构突变的检测。《美国统计学会杂志》**110,** 654–668。

Siegmund, D.（2013）。《序贯分析：检验与置信区间》。Springer Science & Business Media。

Siegmund, D., Yakir, B. 和 Zhang, N. R.（2011）。对齐序列中同时变异区间的检测。《应用统计年鉴》第645–668页。

Siegmund, D. O.（1992）。随机场最大值的尾部近似。载于概率论：*1989*年新加坡概率会议论文集（会议于*1989*年*6*月*8*日至*16*日在新加坡国立大学举行），第147页。Walterde Gruyter GmbH & Co KG。

Torous, J., Kiang,M. V., Lorme, J. 和Onnela,J.-P.（2016）。精神病学新研究的新工具：一个可扩展且可定制的平台，助力数据驱动的智能手机研究。*JMIR* 心理健康**3,** 。

Torous, J., Onnela, J. 和 Keshavan, M.（2017）。实现 RDoC 潜力的新维度与新工具：通过智能手机和互联设备进行数字表型分析。转化精神病学**7,** e1053。

伍德罗夫，M.（1982）。序贯分析中的非线性更新理论。SIAM。

Xie, Y.、Huang, J. 和 Willett, R.（2013）。缺失数据高维时间序列的变点检测。*IEEE* 信号处理精选主题期刊**7,** 12–27。

Zhang, N. R.、Siegmund, D. O.、Ji, H. 和 Li, J. Z.（2010）。检测多个序列中的同步变点。*Biometrika* **97,** 631–645。

收到日期：*2018*年*5*月。修订日期：*??? ????*。接收日期：*??? ????*。

Appendix

定理 *1* 的证明

证明。对于 $k \geqslant n - m_1$, $U_{ik} = \sum_{j=k+1}^{n} Y_{ij}/\sqrt{n} - k + o(1)$。令 P_δ 表示 $\sqrt{k}(U_{1k}$、.....、$U_{qk})^T$ 的概率测度，其中 $E(Y_{ij}) = \delta_i$。我们采用类似渐近尾概率估计中常用的方法，从似然比恒等式入手（Nardi 等，2008）。 P_0 相对于 $A = \prod_{i=1}^{q} \int_{-\infty}^{\infty} P_{\delta_i}\phi(\delta_i)d\delta_i$ 的似然比为

$$\begin{aligned} l(k; A, P_0) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{i=1}^{q} \frac{\frac{1}{\sqrt{2\pi k}} \exp\left\{-\left(\sum_{j=k+1}^{n} Y_{ij} - (n-k)\delta_i\right)^2/2(n-k)\right\}}{\frac{1}{\sqrt{2\pi k}} \exp\left\{-\left(\sum_{j=k+1}^{n} Y_{ij}\right)^2/2(n-k)\right\}} \phi(\delta_i)d\delta_i \\ &= (n-k+1)^{-q/2} \exp\{\textstyle\sum_{i=1}^{q}(\sum_{j=k+1}^{n} Y_{ij})^2/2(n-k+1)\} \end{aligned}$$

令 $K = \max(k; n - m_1 \leqslant k \leqslant n - m_0,\ Z_k \geqslant b^2)$，Wald 的似然比恒等式给出

$$\begin{aligned} \mathrm{pr}(Q \geqslant b^2; P_0) &= E_A\left\{l(K; P_0, A) I_{\{Q \geqslant b\}}\right\} \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (n-K+1)^{q/2} \exp\{-\textstyle\sum_{i=1}^{q}(\sum_{j=K+1}^{n} Y_{ij})^2/2(n-K+1)\} \\ &\quad \cdot I_{\{Q \geqslant b^2\}} \textstyle\prod_{i=1}^{q} \phi(\delta_i)d\delta_i \end{aligned}$$

由于 K 是 Z_k 越过 b^2 的索引，我们有 $Z_K = b^2 + o(1)$，根据 Siegmund（2013）中的 4.12 进行推广后，这意味着 $\mathrm{pr}\{\lim_{b\to\infty}(n-K)b^{-2} = (\sum_{i=1}^{q} \delta_i^2)^{-1}\}$。因此， $Q \geqslant b^2$ 与 $c_1 \leqslant \theta \leqslant c_0$ 渐近等价，其中 $\theta = (\sum_{i=1}^{q} \delta_i^2)^{-1/2}$。

$$\begin{aligned} \mathrm{pr}(Q \geqslant b^2; P_0) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (n-K+1)^{q/2} \exp\{-(n-K)b^2/2(n-K+1)\} \\ &\quad \cdot I_{\{c_1 \leqslant \theta \leqslant c_0\}} \textstyle\prod_{i=1}^{q} \phi(\delta_i)d\delta_i \\ &= (2\pi)^{-q/2} b^q exp(-b^2/2) \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \theta^{-q} I_{\{c_1 \leqslant \theta \leqslant c_0\}} d\delta_1 \cdots d\delta_q \\ &= (2\pi)^{-q/2} b^q exp(-b^2/2) \int_{c_1}^{c_0} \theta^{-q} \left\{\frac{2\pi^{q/2}\theta^{q-1}}{\Gamma(q/2)}\right\} d\theta \end{aligned}$$

最后这个等式是注意到 q维积分是在曲面上进行的。

球面的面积在 $\mathbb{R}^q$中，或 $(2\pi^{q/2}\theta^{q-1})/\Gamma(q/2)$，半径范围从 $c_1 \leqslant \theta \leqslant c_0$。注意到 $\int_{c_1}^{c_0} \theta^{-1} d\theta = 2\ln(c_0/c_1) = \ln(m_1/m_0)$ 导致 4。

Proof of Theorem 2

证明。将 $f\left(\hat{\mathbf{\Sigma}}\right)$ 在对称矩阵 $\mathbf{\Sigma}$ 附近展开，可得:

$$\sqrt{B} f\left(\hat{\mathbf{\Sigma}}\right) = \sqrt{B} f(\mathbf{\Sigma}) + \sqrt{B} \sum_{i<j} \frac{\delta f(\mathbf{\Sigma})}{\delta \Sigma_{ij}} (\hat{\Sigma}_{ij} - \Sigma_{ij}) + o_B(1).$$

对 $f(\mathbf{\Sigma})$ 关于 Σ_{ij} 求导，得到

$$\begin{aligned} \frac{\delta f(\mathbf{\Sigma})}{\delta \Sigma_{ij}} = & -\tfrac{1}{2}|\mathbf{\Sigma}|^{-3/2}(2\pi)^{-m^*/2} \tfrac{\delta|\mathbf{\Sigma}|}{\delta \Sigma_{ij}} \int_{-\infty}^{a} \cdots \int_{-\infty}^{a} \exp\left(-\tfrac{1}{2}\boldsymbol{y}^T \mathbf{\Sigma}^{-1} \boldsymbol{y}\right) d\boldsymbol{y} \\ & + |\mathbf{\Sigma}|^{-1/2}(2\pi)^{-m^*/2} \int_{-\infty}^{a} \cdots \int_{-\infty}^{a} \tfrac{\delta(-\frac{1}{2}\boldsymbol{y}^T \mathbf{\Sigma}^{-1} \boldsymbol{y})}{\delta \Sigma_{ij}} \exp\left(-\tfrac{1}{2}\boldsymbol{y}^T \mathbf{\Sigma}^{-1} \boldsymbol{y}\right) d\boldsymbol{y}. \end{aligned}$$

我们使用 Jacobi 公式，利用 $\delta|\mathbf{\Sigma}|/\delta\Sigma_{ij} = 2|\mathbf{\Sigma}|(\mathbf{\Sigma}^{-1})_{ij}$ 简化第一项。注意到 $\underline{\frac{\delta \mathbf{\Sigma}^{-1}}{\delta \Sigma_{ij}}} = \mathbf{\Sigma}^{-1} \underline{\frac{\delta \mathbf{\Sigma}}{\delta \Sigma_{ij}} \mathbf{\Sigma}^{-1}}$，可得:

$$\begin{aligned} \frac{\delta f(\mathbf{\Sigma})}{\delta \Sigma_{ij}} &= -\left(\mathbf{\Sigma}^{-1}\right)_{ij} f(\mathbf{\Sigma}) + |\mathbf{\Sigma}|^{-1/2}(2\pi)^{-m^*/2} \int_{-\infty}^{a} \cdots \int_{-\infty}^{a} \tfrac{\delta\left(-\frac{1}{2}\boldsymbol{y}^T \mathbf{\Sigma}^{-1} \boldsymbol{y}\right)}{\delta \Sigma_{ij}} \exp\left(-\tfrac{1}{2}\boldsymbol{y}^T \mathbf{\Sigma}^{-1} \boldsymbol{y}\right) d\boldsymbol{y} \\ &= -\left(\mathbf{\Sigma}^{-1}\right)_{ij} f(\mathbf{\Sigma}) + E\left(\tfrac{1}{2}\boldsymbol{y}^T \mathbf{\Sigma}^{-1} \tfrac{\delta \mathbf{\Sigma}}{\delta \Sigma_{ij}} \mathbf{\Sigma}^{-1} \boldsymbol{y}\right) - \{1 - f(\mathbf{\Sigma})\} h_{ij}(a) \\ &= -\left(\mathbf{\Sigma}^{-1}\right)_{ij} f(\mathbf{\Sigma}) + \tfrac{1}{2}\mathrm{tr}\left(\mathbf{\Sigma}^{-1} \tfrac{\delta \mathbf{\Sigma}}{\delta \Sigma_{ij}} \mathbf{\Sigma}^{-1} \mathbf{\Sigma}\right) - \{1 - f(\mathbf{\Sigma})\} h_{ij}(a) \\ &= -\left(\mathbf{\Sigma}^{-1}\right)_{ij} f(\mathbf{\Sigma}) + \left(\mathbf{\Sigma}^{-1}\right)_{ij} - \{1 - f(\mathbf{\Sigma})\} h_{ij}(a) \\ &= \{1 - f(\mathbf{\Sigma})\} \left\{\left(\mathbf{\Sigma}^{-1}\right)_{ij} - h_{ij}(a)\right\} \end{aligned}$$

当 B 较大时，根据 Hotelling (1953)，有 $\mathrm{var}(\hat{\Sigma}_{ij}) = (1 - \Sigma_{ij}^2)^2/B$。我们可以使用一阶展开得到 $f\left(\hat{\mathbf{\Sigma}}\right)$ 的方差:

$$\mathrm{var}\left\{\sqrt{B} f\left(\hat{\mathbf{\Sigma}}\right)\right\} = \{1 - f(\mathbf{\Sigma})\}^2 \sum_{i<j} \left[\left\{\left(\mathbf{\Sigma}^{-1}\right)_{ij} - h_{ij}(a)\right\}\left(1 - \Sigma_{ij}^2\right)\right]^2 + o_B(1)$$

推论 *3* 的证明

证明。首先，注意到 $p < .5$ 意味着 $a > 0$ 对任意 $\mathbf{\Sigma}$。由于 $h_{ij}(a)$是二次型的条件均值，存在 $c^\dagger$ 使得 $h_{ij}(a) \leqslant c^\dagger a^2$ 对所有 $i \neq j$ 以及对所有$a > 0$。由于 c_a 关于 $h_{ij}(a)$ 是二次的，存在 $c^\ddagger$ 使得 $c_a \leqslant c^\ddagger a^4$ 对 $a > 0$。最后，由于

$p = \text{pr}\{\cup_k(Z_k^* > a)\} > 1-\Phi(a) \geqslant \exp(-a^2/2)/\sqrt{2\pi}(\frac{1}{a} - \frac{1}{a^3})$，存在 c^* 使得 $c^\ddagger a^4 \leqslant c^*(\log\ p)^2$对所有 $p < 0.5$。结合公式6和 $\text{var}(\tilde{p}) = p(1-p)/B$，这就完成了证明。

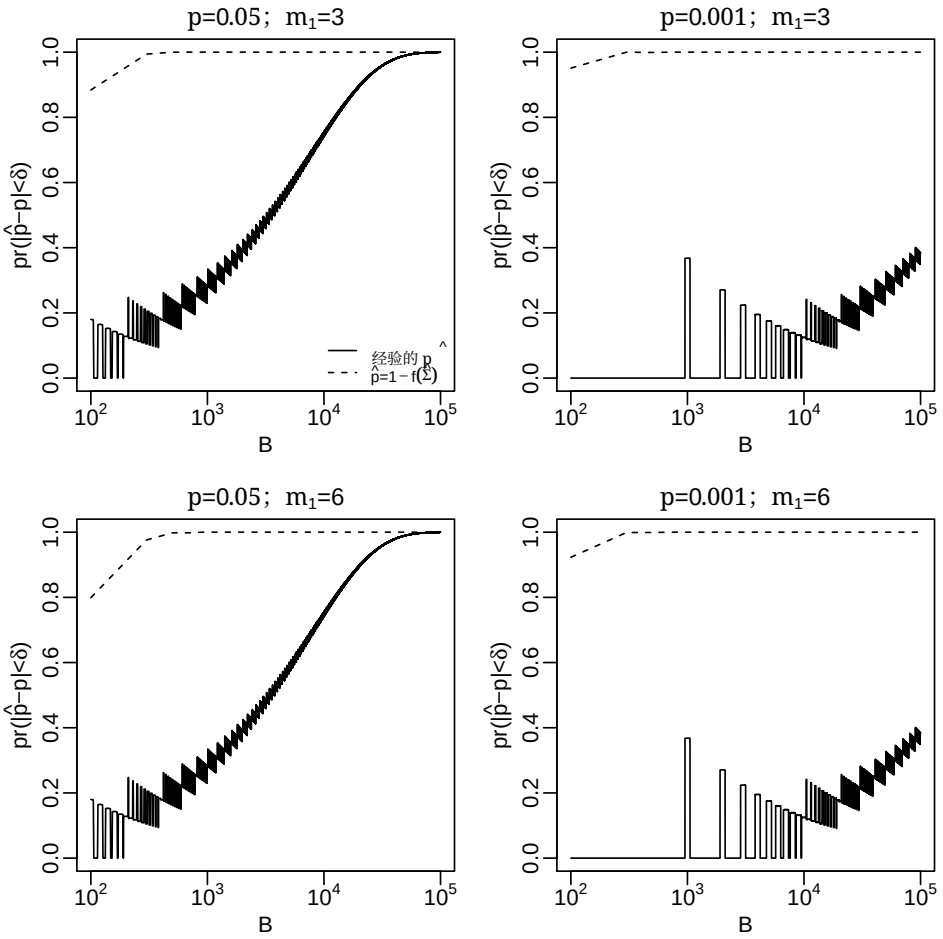


图 1。 估计 p 值的准确性。将 p、 $\tilde{p}$ 的二项比例经验估计（实线）与通过使用 $\boldsymbol{\Sigma}$ 的经验估计量 $\hat{}p$ 对多元正态密度进行积分而获得的 p 估计值（虚线）进行比较。在这两种情况下，均使用 H_0 下的 B 重复模拟来估计 p。准确性由估计 p值中处于 p 的 5% 以内的比例决定（即 $\delta = 0.05p$）。在所有情况下， $m_0 = 0$ 均表示在线变点检测场景。

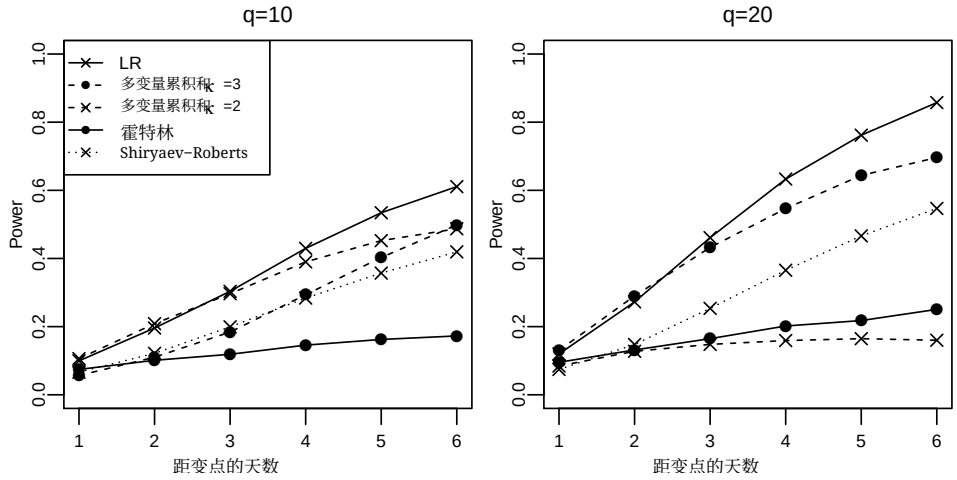


图 2. 多元变点检测方法的功效比较。当变点位置 k 从 $n-1$ 到 $n-6$ 变化时，其中 $n = 100$，此模拟对应于对一名个体进行 100 天数字表型分析监测后的变点检测，变点出现在过去一周的不同时间点。特征根据模型 1 独立生成，每个 $\delta_i = 0.5$。

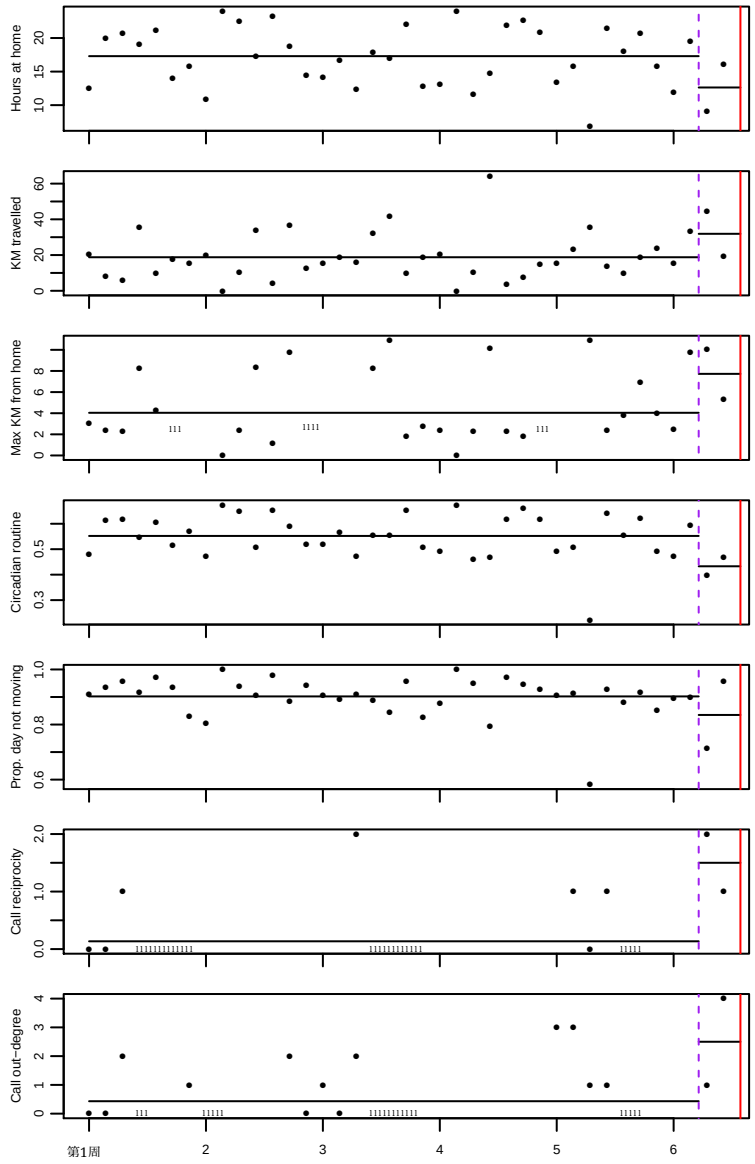


图 3. 精神分裂症患者复发前行为特征的变化。垂直虚线表示变点的时间，垂直实线表示复发并住院的时间。水平实线分别表示变点前后各项特征的均值。

表 1

Type I error simulations for likelihood ratio tests. The asymptotic p-values are generated according to equation 4. The other two approaches obtain p-values using the $\boldsymbol{Z}^*$ with its covariance $\boldsymbol{\Sigma}$ estimated empirically or through a 1st order approximation. The 0.05 *significance level is used with* 10000 *simulations for each setting.*

		$m_1 = 6$ $m_0 = 0$			$m_1 = \sqrt{n}$ $m_0 = \sqrt{n}/2$		
q	n	渐近	近似 $\hat{\boldsymbol{\Sigma}}$	Emp. $\hat{\boldsymbol{\Sigma}}$	渐近	近似 $\hat{\boldsymbol{\Sigma}}$	Emp. $\hat{\boldsymbol{\Sigma}}$
5	30	0.136	0.034	0.044	0.082	0.051	0.039
5	100	0.133	0.036	0.041	0.070	0.089	0.037
5	365	0.144	0.035	0.041	0.068	0.141	0.039
5	1000	0.141	0.042	0.045	0.055	0.173	0.035
10	30	0.169	0.030	0.042	0.095	0.070	0.044
10	100	0.175	0.041	0.045	0.085	0.097	0.040
10	365	0.340	0.048	0.048	0.077	0.150	0.042
10	1000	0.163	0.040	0.042	0.069	0.185	0.043
50	30	0.332	0.038	0.045	0.197	0.079	0.050
50	100	0.329	0.049	0.050	0.169	0.101	0.047
50	365	0.340	0.048	0.049	0.154	0.151	0.047
50	1000	0.340	0.050	0.048	0.138	0.188	0.047