Advances in Applied Mathematics 应用数学进展, 2023, 12(2), 537-549
Published Online February 2023 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2023.122057

Hans汉斯

聚类复发事件加性模型的分段常数化估计

郭　蒋

北京信息科技大学理学院，北京

收稿日期：2023年1月16日；录用日期：2023年2月11日；发布日期：2023年2月21日

摘　要

研究大型的医疗数据具有重要意义，但大量数据会导致计算复杂度变高，尤其涉及到反复发生的疾病数据时。本文对大型数据的聚类复发事件提出了一个半参数边际加性比率函数模型。在进行参数估计时，随着数据量级的增加而导致未知参数增加，为了提高计算效率，对基准比率函数进行分段常数化。利用估计方程方法给出了模型参数的估计量的表达式。并证明估计量满足相合性和渐近正态性等大样本性质。通过数值模拟，验证了提出的估计方法，模拟结果表明对参数部分的估计结果较好。与对基准比率函数未做分段常数化的方法进行比较，发现分段常数化的方法偏差更小，而且计算时间显著缩短。最后将模型和方法应用到慢性肉芽肿病的数据中，将病人按照医院进行分类，找到影响各医院病人病情复发的显著因素和因素的影响方式。

关键词

大型医疗数据，聚类复发事件，边际比率模型，估计方程，分段常数化

Estimation of an Additive Rate Model for Clustered Recurrent Events Based on the Piecewise Constant Method

Jiang Guo

College of Science, Beijing Information Science & Technology University, Beijing

Received: Jan. 16th, 2023; accepted: Feb. 11th, 2023; published: Feb. 21st, 2023

Abstract

Studying large medical databases is important. However, the use of large databases may introduce computational difficulties, particularly when the event of interest is recurrent. A semiparametric marginal additive rate function model is proposed for large databases clustered recurrence events.

文章引用: 郭蒋. 聚类复发事件加性模型的分段常数化估计[J]. 应用数学进展, 2023, 12(2): 537-549.
DOI: 10.12677/aam.2023.122057

In the parameter estimation, the number of baseline rate functions increases as the number of classes increases, leading to an increase in the parameters. The baseline rate functions were regarded as piecewise constant during the estimation, which improved the computational efficiency. The expressions of the estimate of the model parameters were given using the estimating equation method. The estimators were proved consistent and asymptotic normality. The proposed estimation method is verified by numerical simulation, and the simulation results show that the estimation results of the parameter part are good, compared with the method without piecewise constant of the risk ratio function. It is found that the bias of the piecewise constant method is smaller, and the calculation time is significantly shortened. Finally, the model and method were applied to the data of chronic granulomatous disease, and the patients were classified by hospital to find out the significant factors affecting patients' disease recurrence in each hospital and the influencing ways of the factors.

Keywords

Large Medical Databases, Clustered Recurrent Events, Marginal Rate Model, Estimating Equation, Piecewise Constant

1. 引言

在一段时间内，若某个感兴趣事件重复发生，我们称之为复发事件。复发事件在临床试验以及工业领域中很常见，例如癌症的反复发生，反复感染病毒，机器反复发生故障等等。在实际研究中，我们遇到的复发数据可能是来自不同医疗机构的病人，或者是家族遗传病，这时，来自于同一个机构或家庭的病人就具有相似性，我们需要根据这些属性(例如医疗机构，家族)，将研究的多个个体进聚类，得到聚类复发事件数据。

复发事件是近些年来的一个热门研究领域。根据复发事件建模对象的不同，主要有两种研究：一种是对复发事件过程的强度函数建模[1] [2]；另一种是对边际均值或比率函数进行建模[3]。均值模型和比率模型具有计算简单，真值易估计、稳健性强等优点，因此吸引了大批学者进行研究。有些学者假设协变量效应和基准比率之间是乘积关系的乘性模型，例如 Lin [4]或者 Lawless [5]。相应的，也有学者研究协变量效应和基准比率之间是相加关系的加性模型，例如 Liu 和 Wu [6]等人。Cai [7]等人则在平均寿命函数中同时考虑了加性效应和乘性效应。死亡事件也受到了学者的关注，Sun [8]等人提出了一种特定复发事件均值的半参数方法，该方法结合了死亡事件的加性模型和条件复发事件的加性模型，在这篇论文的基础上，马燕[9]所提出的方法则是基于乘性模型。Xu [10]等人还研究了复发事件和死亡事件的联合尺度变化模型，并用脆弱变量来解释复发事件和死亡事件之间的相关性，Chan 和 Wang [11]通过以死亡时间为源头，通过逆向计数的方法来研究复发事件。在估计间歇性观测的时间依赖协变量的比率模型时，Sun 等人在 2020 [12]年提出了一种基于逆速率加权和核平滑的估计方法，Lyu Tianmeng 等人 2021 [13]年则提出了使用加性－乘性比率的半参数回归模型来解决上诉问题。对被调查的个体可能会经历几种不同类型的多类型复发事件，杜彦斌[14]等人提出了一种半参数转移模型，作者利用广义估计方程的思想对参数进行了估计。杨青龙[15]用广义半参数风险模型研究了多类型复发事件，并用重采样方法计算估计参数的方差。

对于聚类复发事件的研究主要集中在边际均值或者边际比率函数上。Schaubel 和 Cai [16]对聚类复发

事件提出了一个半参数乘积比率模型。Liu [17]推广了该模型，假设来自于一个中心的数据为一个类，不同的类具有不同的基准比率函数，并且考虑了死亡事件，当中心数较大时，作者提出对基准比率函数分段常数化，提高计算效率，但是 Liu [17]只研究了乘积比率模型，没有考虑加性的情况。Liu 等人[18]在聚类复发事件的比率函数加入了一个起乘积作用的类固定效应项，但该方法的计算效率不高。He [19]等人提出了一个加性比率函数模型，在模型中加入带加性作用的类固定效应项。此外，Sun [20]等人还分别考虑了加性比率模型和加性风险回归模型的核加权估计过程。

本论文借鉴了 Liu [17]等人的基准比率函数分段常数化的方法，但他只研究了乘性比率函数模型的情况，我们在此基础上研究加性比率函数模型的估计，并且进一步研究了估计量大样本性质。

本文第一部分给出记算符号，并建立模型和估计方法，第二部分给出估计量的相合性和渐近正态性等大样本理论，第三部分进行数值模拟验证提出的方法，并与基准比率函数不做分段化处理时的结果进行了比较。第四部分分析了一个慢性肉芽肿病的数据，第五部分是论文的总结和展望，渐近性质的证明在附录里。

2. 模型和估计方法

2.1. 符号说明及模型建立

假设有 n 个观测的个体，将这些研究的个体分成 K 个类，用 $k\left(k=1,2,\cdots,K\right)$ 表示类的指标值，用 n_k 代表第 k 个类所包含的样本数，满足 $\sum_{k=1}^{K} n_k = n$，同时用 $i\left(i=1,2,\cdots,n\right)$ 表示第 i 个个体，对于个体 i，我们用 G_i 表示该个体的类别，$G_i = k$ 表示个体 i 在第 k 个类。C_i 表示个体 i 的右删失时间，假设删失时间与复发事件过程是独立的。$Y_i(t) = I\left(t \le C_i\right)$ 表示个体处于删失的风险，这里的 $I(\cdot)$ 是一个示性函数。用 $N_i^*(t)$ 表示在 t 时刻之前，个体 i 潜在的事件复发次数，$N_i(t) = \int_0^t Y_i(s)\mathrm{d}N_i^*(s)$ 表示在 t 时刻之前观测到的复发次数。用 τ 表示观测截止时间。假设给定协变量的条件下，右删失时间与复发事件过程是独立的。

用 $\mathrm{d}N_i^*(t) = N_i^*(t) - N_i^*(t-)$ 表示第 i 个个体在时刻 t 的瞬间复发事件发生次数，它只能取 1 (有事件发生)或 0 (没有事件发生)。在复发事件数据分析中，定义 $E\left\{\mathrm{d}N_i^*(t) \mid N_i^*(u), Z_i(u), 0 \le u < t, G_i = k\right\}$ 为给定历史数据下，t 时刻复发事件的风险或强度函数。我们研究复发事件的边际强度函数 $E\left\{\mathrm{d}N_i^*(t) \mid Z_i(t), G_i = k\right\}$，也叫做比率函数。对比率函数定义如下的加性模型：

$$E\left\{\mathrm{d}N_i^*(t) \mid Z_i(t), G_i = k\right\} = \left(\rho_{0k}(t) + Z_i(t)^{\mathrm{T}}\beta\right)\mathrm{d}t, \tag{1}$$

式中：$Z_i(t)$ 表示与时间有关的协变量，$\rho_{0k}(t)$ 表示与类别有关的基准比率函数，β 是待估计的参数。

定义 $G_{ik} = I\left(G_i = k\right)$，$Y_{ik}(t) = G_{ik}Y_i(t)$，$\mathrm{d}N_{ik}^*(t) = G_{ik}\mathrm{d}N_i^*(t)$，$\mathrm{d}N_{ik}(t) = G_{ik}\mathrm{d}N_i(t)$，在右删失与复发事件独立的假设下，有 $E\left\{\mathrm{d}N_i^*(t) \mid Z_i(t), G_i = k, Y_i(t) = 1\right\} = E\left\{\mathrm{d}N_i(t) \mid Z_i(t), G_i = k\right\}$，结合 $N_i(t) = \int_0^t Y_i(s)\mathrm{d}N_i^*(s)$，模型(1)变成

$$E\left\{\mathrm{d}N_{ik}(t) \mid Z_i(t), Y_{ik}(t)\right\} = Y_{ik}(t)\left(\rho_{0k}(t) + Z_i(t)^{\mathrm{T}}\beta\right)\mathrm{d}t. \tag{2}$$

2.2. 估计方法

我们利用分段常数化的思想，将 $[0,\tau]$ 切割成 L 个区间，$a_0 < a_1 < \cdots < a_L$ 是 L 个区间的切割点。设 $\rho_{0k}(t)$ 在区间 $\left(a_{l-1}, a_l\right]$ 上为一个常值函数，记成 ρ_{kl}。同样的，将 $z_i(t)$ 进行分段常值化，用 z_{il} 表示 $z_i(t)$ 在时间区间 $\left(a_{l-1}, a_l\right]$ 上的值。记 $d_{ikl} = \int_{a_{l-1}}^{a_l} \mathrm{d}N_{ik}(t)$，$t_{ikl} = \int_{a_{l-1}}^{a_l} Y_{ik}(t)\mathrm{d}t$，根据以上的记法，模型(2)以及估计方程的思想，得到如下的估计方程：

$$\sum_{i=1}^{n}\left[d_{ikl}-t_{ikl}\left(\rho_{kl}+\beta^{\mathrm{T}}Z_{il}\right)\right]=0, l=1,2,\cdots,L \tag{3}$$

$$\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{L}\int_{a_{l-1}}^{a_l}Z_{il}\left[\mathrm{d}N_{ik}\left(t\right)-Y_{ik}\left(t\right)\cdot\left(\rho_{kl}+Z_{il}^{\mathrm{T}}\beta\right)\mathrm{d}t\right]=0 \tag{4}$$

由方程(3)可得

$$\rho_{kl}=\frac{\sum_{i=1}^{n}d_{ikl}-\sum_{i=1}^{n}t_{ikl}\beta^{\mathrm{T}}Z_{il}}{\sum_{i=1}^{n}t_{ikl}}, l=1,2,\cdots,L \tag{5}$$

将(5)式代入方程(4)，可得

$$\hat{\beta}=\left[\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{L}t_{ikl}\left(Z_{il}-\bar{Z}_{kl}\right)^{\otimes 2}\right]^{-1}\cdot\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{L}\left(Z_{il}-\bar{Z}_{kl}\right)\cdot d_{ikl}, \tag{6}$$

其中 $\bar{Z}_{kl}=\frac{\sum_{i=1}^{n}t_{ikl}Z_{il}}{\sum_{i=1}^{n}t_{ikl}}$。在上式中，对于每一个向量 v，$v^{\otimes 2}=vv'$。将 β 的估计量，代入到(5)式，得到 ρ_{kl} 的估计结果如下：

$$\hat{\rho}_{kl}\left(\hat{\beta}\right)=\frac{\sum_{i=1}^{n}d_{ikl}-\sum_{i=1}^{n}t_{ikl}\hat{\beta}^{\mathrm{T}}Z_{il}}{\sum_{i=1}^{n}t_{ikl}}, \tag{7}$$

根据分段常值化的假定，我们可以进一步求出累积基准比率函数 $\mu_{0k}\left(t\right)=\int_0^t\rho_{0k}\left(u\right)\mathrm{d}u$ 的估计：

$$\hat{\mu}_{0k}\left(t;\hat{\beta}\right)=\sum_{l=1}^{L}\hat{\rho}_{kl}\left(\hat{\beta}\right)\left(\alpha_l\wedge t-\alpha_{l-1}\wedge t\right) \tag{8}$$

式中：$a\wedge b=\min\left(a,b\right)$，表示两个数中的较小者。

3. 渐近性质

记 $\bar{z}_{kl}=\frac{E\left(t_{ikl}Z_{il}\right)}{E\left(t_{ikl}\right)}$ 是 $\bar{Z}_{kl}=\frac{\sum_{i=1}^{n}t_{ikl}Z_{il}}{\sum_{i=1}^{n}t_{ikl}}$ 的极限，$\bar{t}_{kl}=E\left(t_{ikl}\right)$ 是 $\frac{1}{n}\sum_{i=1}^{n}t_{ikl}$ 的极限。另外记 $\mathrm{d}M_{ikl}=d_{ikl}-t_{ikl}\left(\rho_{kl}+Z_{il}'\beta_0\right)$。

为了研究估计量的渐近性质，我们需要以下正则性条件：

a) $P\left(C_{ij}\geq\tau\right)>0$ 对于一切 $\tau>0$ 都成立；

b) $E\left(t_{ikl}\right)>0$，对于所有的 $l=1,2,\cdots,L$ 均成立。(为了保证 $\bar{z}_{kl}$ 的分母不等于 0)；

c) $d_{ikl},Z_{il},l=1,\cdots,L$，都是有界的；

d) A 是正定矩阵，这里 $A=E\left[\sum_{k=1}^{K}\sum_{l=1}^{L}t_{ikl}\left(Z_{il}-\bar{z}_{kl}\right)^{\otimes 2}\right]$。

3.1. 定理一

在上述正则化条件下，$\hat{\beta}$ 是一个强相合估计。收敛关系 $\hat{\beta}\to\beta_0$ 几乎处处成立。而且 $\sqrt{n}\left(\hat{\beta}-\beta_0\right)$ 收

敛到一个正态分布，这个正态分布的均值是 0，方差是 $A^{-1}\Omega A^{-1}$，可以由 $\hat{A}^{-1}\hat{\Omega}\hat{A}^{-1}$ 估计，这里 $\eta_i=\sum_{k=1}^{K}\sum_{l=1}^{L}\left(Z_{il}-\bar{z}_{kl}\right)\mathrm{d}M_{ikl}$。

3.2. 定理二

在上述条件下，对于任意给定的 k，$k=1,2,\cdots,K$，$\hat{\mu}_{0k}\left(t,\hat{\beta}\right)$关于$t\in(0,\tau]$几乎处处一致收敛于 $\mu_0(t)$，$\sqrt{n}\left[\hat{\mu}_{0k}\left(t,\hat{\beta}\right)-\mu_0(t)\right]$弱收敛于零均值的高斯过程，其在$(s,t)$处的协方差函数为 $\Gamma_k(s,t)=\sum_{j=1}^{n}E\left(\phi_{jk}(s)\phi_{jk}(t)\right)$，这里

$$\hat{\phi}_{jk}(t)=\left(-\frac{1}{\sqrt{n}}\sum_{l=1}^{L}\bar{z}_{kl}'\left(A^{-1}\eta_j\right)\right)\left(\alpha_l\wedge t-\alpha_{l-1}\wedge t\right)+\sqrt{n}\sum_{l=1}^{L}\frac{\mathrm{d}M_{jkl}}{\bar{t}_{kl}}\left(\alpha_l\wedge t-\alpha_{l-1}\wedge t\right)$$

4. 数值模拟

4.1. 数据生成

在本节中，我们按照模型随机生成有限的样本，对参数进行模拟研究。首先，我们假设 n 个个体来自于 K 个类($K=10,50$)，每个类包含 n_k 个样本($n_k=2,5,10,20$)，总个体 $n=K\times n_k$，对每一个个体 $i\left(i=1,2,\cdots,n\right)$我们生成两个协变量 Z_{1i},Z_{2i}，对于第 k 类生成一个脆弱变量 $W_k\left(k=1,2,\cdots,K\right)$，假设 Z_{1i} 来自区间(0, 1)上的均匀分布，Z_{2i} 来自参数为 0.5 的两点分布，W_k 来自于均值和方差都为 1 的伽玛分布，并假设来自同一个类的个体共用一个相同的 W_k。假设删失时间 C 服从区间$(u, v]$上的均匀分布，u 在[0, 0.5]之间变化，v 在[2, 6]之间变化，试验终止时间 $\tau=3$，这些设置保证复发事件的平均发生次数在 2~3 次之间。我们从下列模型中生成复发事件：

$$E\left\{\mathrm{d}N_i^*(t)\mid Z_i(t),G_i=k,W_k\right\}=\left(\rho_{0k}(t)+Z_i(t)^{\mathrm{T}}\beta_0+W_k\right)\mathrm{d}t \tag{9}$$

我们分别考虑 $\rho_{0k}(t)=1$、$\rho_{0k}(t)=t$ 两种情况，$\beta_0=(0.5,0.5)'$，为了对基准比率函数实行分段常数化，将事件发生时间的整个区间$(0,\tau]$分别分成 $L=10$ 或 $L=15$ 个小区间，在每一个小区间上 $\rho_{0k}(t)$ 为一个常数。我们根据提出的方法来估计模型中的回归参数 β 和累积基准比率函数 $\mu_{0k}(t)=\int_0^t\rho_{0k}(u)\mathrm{d}u$。

4.2. 模拟结果

β 的估计结果见表1和见表2。表中BIAS表示 β 的估计值 $\hat{\beta}$ 和真值 β_0 之间的偏差在1000次中的平均值，ASE表示 $\hat{\beta}$ 的计算的标准差估计在1000次模拟中的平均值，ASD表示1000次模拟中所得到的 $\hat{\beta}$ 的观测值的样本标准差，CP表示在正态分布下，真值 β_0 的95%置信区间内，覆盖到的 $\hat{\beta}$ 的百分比。模拟结果可以看出：估计值和真值的偏差都在0.05以内，样本标准差和估计量的标准差估计比较接近，β_0 的95%置信区间内，覆盖真值的比例在95%附近，估计效果较好。

Table 1. The parameter estimation results of repeated simulations 1000 times when $\mu_{0k}=t$
表 1. $\mu_{0k}=t$ 时，重复模拟 1000 次的参数估计结果

			β_1				β_2			
K	n_k	L	BIAS	ASE	ASD	CP	BIAS	ASE	ASD	CP
10	10	10	−0.004	0.431	0.468	0.919	0.014	0.249	0.269	0.934
	10	15	0.013	0.430	0.452	0.921	0.010	0.250	0.263	0.930

Continued

10	20	10	0.004	0.306	0.311	0.938	0.015	0.176	0.181	0.953
	20	15	0.009	0.306	0.314	0.942	0.014	0.176	0.188	0.930
50	5	10	0.021	0.272	0.311	0.915	0.012	0.157	0.181	0.910
	5	15	0.026	0.274	0.306	0.913	0.009	0.157	0.184	0.911
50	10	10	0.013	0.192	0.210	0.925	0.009	0.111	0.121	0.934
	10	15	0.006	0.164	0.175	0.931	0.002	0.095	0.102	0.925

Table 2. The parameter estimation results of repeated simulations 1000 times when $\mu_{0k} = 0.5t^2$

表 2. $\mu_{0k} = 0.5t^2$ 时，重复模拟 1000 次的参数估计结果

			β_1				β_2			
K	n_k	L	BIAS	ASE	ASD	CP	BIAS	ASE	ASD	CP
10	10	10	0.011	0.429	0.471	0.933	0.017	0.249	0.268	0.929
	10	15	0.002	0.429	0.454	0.935	0.031	0.248	0.269	0.915
10	20	10	0.007	0.304	0.315	0.934	0.020	0.176	0.183	0.945
	20	15	0.022	0.307	0.307	0.933	0.016	0.177	0.177	0.937
50	5	10	−0.007	0.243	0.272	0.921	0.004	0.140	0.160	0.910
	5	15	−0.012	0.230	0.259	0.921	0.002	0.133	0.1150	0.921
50	10	10	0.004	0.172	0.190	0.930	−0.001	0.099	0.110	0.923
	10	15	0.008	0.163	0.169	0.938	−0.002	0.094	0.098	0.938

另外，我们也绘制了 $K = 50$， $n_k = 10$， $L = 15$ 时， $\mu_{01}(t)$ 在不同时间点处的估计曲线以及置信度为95%的置信带的估计，见图1和图2。 $\mu_{01}(t)$ 的估计的95%置信带内基本上包含了真实的 $\mu_{01}(t)$ 。图1和图2如下所示。

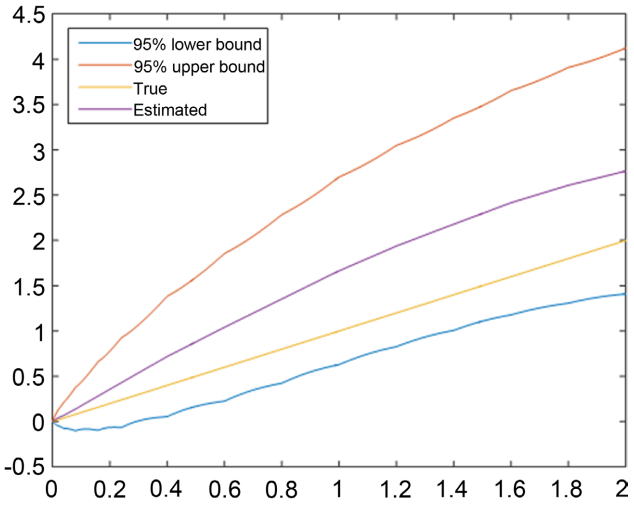

Figure 1. Estimate when $\mu_{01}(t) = t$

图 1. $\mu_{01}(t) = t$ 时的估计

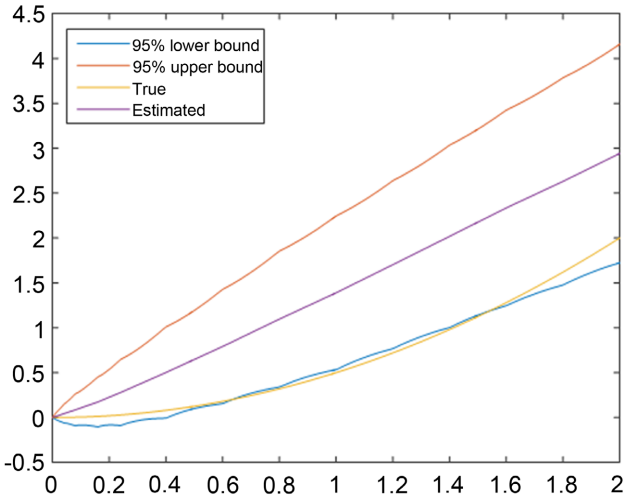

Figure 2. Estimate when $\mu_{01}(t)=0.5t^2$

图 2. $\mu_{01}(t)=0.5t^2$ 时的估计

在评价加入分段常数化的计算效率的时候，我们使用联想YOGA14S笔记本电脑，该电脑基于AMD R7-5800H处理器。在设置模型参数时，我们设置 $n=800$，$\beta=[0.5,0.5]$，$\rho_{0k}(t)=t$，用本文的分段常数化方法与不加入分段常数化的传统方法进行了比较。在不进行分段常数化时，根据计算，β 的估计表达式为：

$$\hat{\beta}=\left(\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^{\tau}Y_{ik}(t)\left(Z_i(t)-\bar{Z}_k(t)\right)^{\otimes 2}\mathrm{d}t\right)^{-1}\cdot\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^{\tau}\left(Z_i(t)-\bar{Z}_k(t)\right)\mathrm{d}N_{ik}(t) \tag{10}$$

其中 $\bar{Z}_k(t)=\dfrac{\sum_{i=1}^{n}Z_i(t)Y_{ik}(t)}{\sum_{i=1}^{n}Y_{ik}(t)}$，相应的 $\bar{z}_k(t)=\dfrac{E\left(Z_i(t)Y_{ik}(t)\right)}{E\left(Y_{ik}(t)\right)}$。

同理，此时也可以计算与标准差 $\hat{\beta}$ 有关的值，分别如下：

$$A=E\left(\int_0^{\tau}\sum_{i=1}^{n}\sum_{k=1}^{K}Y_{ik}(t)\left(Z_i(t)-\bar{z}_k(t)\right)^{\otimes 2}\mathrm{d}t\right) \tag{11}$$

$$\eta_i=\sum_{k=1}^{K}\int_0^{\tau}\left(Z_i(t)-\bar{z}_k(t)\right)\mathrm{d}M_{ik}(t) \tag{12}$$

比较的结果在表3中，从结果中发现使用传统的估计方法，估计量的偏差大于分段常数化的偏差结果。无论使用哪种方法，在估计方差时，样本观测的标准差和模拟计算的结果均十分接近。

两种方法的差别主要体现在计算时间上，对于恒定的个体 $n=800$，分别设置 $K=20,40,80$ 计算 500 次模拟结果，结果记录在表 3 中。可以看到，用分段常数化的方法运行时间显著低于传统方法，传统方法所用的时间是分段常数的 4 倍以上；并且随着 K 的增大，这一比值也在不断增大。由于计算机性能的限制，包含更多个体的实验并没有展示。但分段常数化方法偏差更小，计算速度更快是毋庸置疑的，相较于传统方法优势很明显。

了解释分段常数化计算效率更高的原因，假设所有个体n的复发次数之和为 N_r，当我们使用分段常数化的方法计算文中的 d_{ikl}，t_{ikl} 和 ρ_{kl} 时，我们将时间区间一共分成了L段，这里的L由我们设置，此时我们设置的L一般远小于总个体n的复发次数之和 N_r。但当我们使用传统方法时，我们需要计算所有个体相

邻两个发病时间点的 d_{ikl}，t_{ikl} 和 ρ_{kl}，此时的 $L=N_r+1$。随着L的增大。计算所涉及的矩阵维度也随之增大，计算的数据量同样增大，因此计算时间大幅增加，每次计算都带有一定的误差，随着计算量级的增大，计算次数变多，最后所得到的偏差也随之增大。

Table 3. Computational time comparison results
表 3. 运算时间比较结果

		β_1			β_2			
是否分段	k	BIAS	ASE	ASD	BIAS	ASE	ASD	运算时长/min
否	20	−0.0501	0.132	0.135	−0.0498	0.076	0.085	161.3
是	20	0.0299	0.151	0.166	0.0131	0.087	0.082	53
否	40	−0.0457	0.133	0.138	−0.0386	0.077	0.082	283.6
是	40	−0.0069	0.152	0.158	0.0088	0.088	0.088	59
是	80	0.0048	0.15	0.161	0.0172	0.087	0.097	68.3
否	80	−0.0209	0.132	0.145	−0.046	0.076	0.085	564.9

5. 实例分析

在这部分，我们分析了一个慢性肉芽肿病(CGD)数据[16]，这种疾病是免疫功能方面的遗传性疾病，大多发生在婴幼儿中。数据集包括来自不同医院的(13 个医院)128 个的病人，记录了他们的发病的间隔时间，是否接受 Gamma 干扰素治疗，基因信息，年龄，身高，体重，性别，是否删失，是否用过抗生素等信息。根据以前学者的研究[4]，我们主要关心患者年龄和接受 Gamma 干扰素治疗对疾病的重复感染比率函数的影响。用协变量 Z_{i1} 表示是否使用 Gamma 干扰素，$Z_{i1}=1$ 表示使用 Gamma 干扰素，$Z_{i1}=2$ 表示使用安慰剂。用 Z_{i2} 表示年龄，这里我们对年龄做了规范化变换 $Z_{i2}^{*}=\dfrac{Z_{i2}-\min\limits_{i=1,\cdots,128}\left(Z_{i2}\right)}{\max\limits_{i=1,\cdots,128}\left(Z_{i2}\right)-\min\limits_{i=1,\cdots,128}\left(Z_{i2}\right)}$。假设重复感染的比率函数满足模型：

$$E\left\{\mathrm{d}N_i^{*}(t)\mid Z_i,G_i=k\right\}=\left(\rho_{0k}(t)+Z_i^{\mathrm{T}}\beta\right)\mathrm{d}t \tag{13}$$

应用提出的估计方法对参数 β 进行估计，β 的估计结果见表 4 中。从表四的结果可以看出，协变量 Z_{i1} 的估计系数 β_1 为 0.25%，且该系数的检验 P 值小于 0.0001，说明系数是显著的，因此说明接受 Gamma 干扰素治疗能够显著降低重复感染的比率函数。且固定年龄时，使用 Gamma 干扰素治疗的患者比使用安慰剂的患者平均降低 0.25%。协变量 Z_{i2} 的估计系数 β_2 为−0.14%，但是检验的 P 值却大于 0.05，所以年龄的影响不显著。

Table 4. Parameter estimation of CGD data in model (1)
表 4. CGD 数据在模型(1)下的参数估计

β_1			β_2		
EST	ASE	p_value	EST	ASE	p_value
0.25%	0.0006	0.000	-0.14%	0.0013	0.4989

为了对模型进行检验，我们绘制了残差 $r_{ik}=\int_0^{\tau}\mathrm{d}N_{ik}(t)-Y_{ik}(t)\left(\rho_{0k}(t)+Z_i(t)^{\mathrm{T}}\beta\right)\mathrm{d}t$ 对观测个体序号 i

的残差图，见图 3。在图 3 中，横轴表示个体序号，纵轴表示该个体在模型下的残差值。我们发现残差值在[−2, 2]之间随机变化。因此，没有理由拒绝提出的模型。

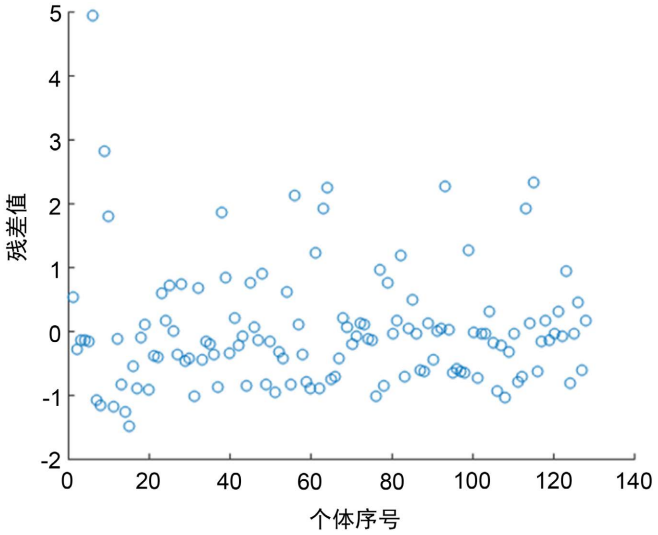

Figure 3. Scatterplot of residuals vs. individuals under additive ratio model
图 3. 加性比率模型下残差对个体的散点图

6. 总结与展望

本文提出了一个聚类复发事件的加性比率模型，该模型将具有相似特征或来源的个体进行了分类，应用估计方程和非参部分分段常数化的方法对参数进行了估计，证明了估计量的相合性和渐近正态性。在数值模拟中获得了较好的估计效果，在与传统方法进行对比时，计算效率远远高于传统方法。并且对慢性肉芽肿病数据进行了分析，得到了与实际情况相吻合的结论。我们提出的模型和方法具有一定的应用前景，特别是可以应用到多中心医疗数据，家族遗传病医疗数据，以及社区医疗数据中，同时，我们的模型计算效率高，处理大规模医疗数据非常得心应手。

属于同一个类别的个体可能具有相关性，未来我们会考虑在模型中加入一个随机效应项，表示同一类个体之间的关系，那就需要对随机效应项的分布做一些假设，对分布中包含的参数进行估计。另外复发事件往往伴随着死亡事件的发生，我们会进一步研究这两类事件之间的联合建模，比如考虑加入随机效应项或者考虑对反向均值函数进行建模。加性模型可以推广到更加灵活的模型上，比如加性乘积模型，变系数模型，转移函数模型，部分线性模型等等，这些模型会用到一些非参数估计方法，统计推断会更复杂一些，更具有挑战性，适用范围也更广。

参考文献

[1] Adersen, P.K. and Gill, R.D. (1982) Cox's Regression Model for Counting Processes: A Large Sample Study. *Annals of Statistics*, **10**, 1100-1120. https://doi.org/10.1214/aos/1176345976

[2] Zeng, D.L. and Lin, D.Y. (2006) Efficient Estimation of Semiparametric Transformation Models for Counting Processes. *Biometrika*, **93**, 627-640. https://doi.org/10.1093/biomet/93.3.627

[3] Lin, D.Y., Wei, L.J., Yang, I. and Ying, Z. (2000) Semiparametric Regression for the Mean and Rate Function of Recurrent Events. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), **62**, 711-730. https://doi.org/10.1111/1467-9868.00259

[4] Dai, J.J., Sun, L.Q. and Yang, Z.H. (2009) A General Additive-Multiplicative Rates Model for Recurrent Event Data. *Science in China*, **52**, 2257-2265. https://doi.org/10.1007/s11425-009-0095-6

[5] Lawless, J.F., Nadeau, C. and Cook, R.J. (1997) Analysis of Mean and Rate Functions for Recurrent Events. In: Lin, D.Y. and Fleming, T.R., Eds., *Proceedings of the First Seattle Symposium in Biostatistics*. *Lecture Notes in Statistics*, Vol. 123, Springer, New York, 37-49. https://doi.org/10.1007/978-1-4684-6316-3_4

[6] Liu, Y.Y. and Wu, Y.S. (2011) Semiparametric Additive Intensity Model with Frailty for Recurrent Events. *Acta Mathematica Sinica*, *English Series*, **27**, 1831-1842. https://doi.org/10.1007/s10114-011-8232-x

[7] Cai, J.H., He, H.J., Song, X.Y. and Sun, L.Q. (2017) An Additive-Multiplicative Mean Residual Life Model for Right-Censored Data. *Biometrical Journal*, **59**, 579-592. https://doi.org/10.1002/bimj.201600068

[8] Sun, X.W. and Sun, L.Q. (2019) Semiparametric Estimation of Differences in Treatment-Specific Recurrent Event Means with a Terminal Event. *Statistics and Its Interface*, **12**, 1-9. https://doi.org/10.4310/SII.2019.v12.n1.a1

[9] 马燕, 孙晓伟. 带终止事件的特定治疗复发事件均值比例的半参估计[J]. 应用数学学报, 2020, 43(6): 949-965.

[10] Xu, G.J, Chiou, S.H., Huang, C.Y. and Wang, M.C. (2017) Joint Scale-Change Models for Recurrent Events and Failure Time. *Journal of the American Statistical Association*, **112**, 794-805. https://doi.org/10.1080/01621459.2016.1173557

[11] Chan, K.C.G. and Wang, M.C. (2017) Semiparametric Modeling and Estimation of the Terminal Behavior of Recurrent Marker Processes Before Failure Events. *Journal of the American Statistical Association*, **112**, 351-362. https://doi.org/10.1080/01621459.2016.1140051

[12] Sun, Y.F., McCulloch, C.E., Marr, K.A. and Huang, C.-Y. (2021) Recurrent Events Analysis with Data Collected at Informative Clinical Visits in Electronic Health Records. *Journal of the American Statistical Association*, **116**, 593-604. https://doi.org/10.1080/01621459.2020.1801447

[13] Lyu, T., Luo, X.H. and Sun, Y.F. (2021) Additive-Multiplicative Rates Model for Recurrent Event Data with Intermittently Observed Time-Dependent Covariates. *Journal of Data Science*, **19**, 615-633. https://doi.org/10.6339/21-JDS1027

[14] 杜彦斌, 戴家佳, 金君. 多类型复发事件数据下一类半参数转移模型[J]. 统计与信息论坛, 2018, 33(4): 20-24.

[15] 杨青龙, 江芹. 多类型复发间隔时间下广义半参数风险模型[J]. 应用数学学报, 2022, 45(1): 72-87.

[16] Schaubel, D.E. and Cai, J. (2005) Semiparametric Methods for Clustered Recurrent Event Data. *Lifetime Data Analysis*, **11**, 405-425. https://doi.org/10.1007/s10985-005-2970-y

[17] Liu, D.D., Schaubel, D.E. and Kalbfleisch, J.D. (2012) Computationally Efficient Marginal Models for Clustered Recurrent Event Data. *Biometrics*, **68**, 637-647. https://doi.org/10.1111/j.1541-0420.2011.01676.x

[18] Liu, D.D., Kalbfleisch, J.D. and Schaubel, D.E. (2014) Methods for Estimating Center Effects on Recurrent Events. *Statistics in Biosciences*, **6**, 19-37. https://doi.org/10.1007/s12561-012-9075-4

[19] He, H.J., Pan, D., Sun, L., *et al*. (2017) Analysis of a Fixed Center Effect Additive Rates Model for Recurrent Event Data. *Computational Statistics & Data Analysis*, **112**, 186-197. https://doi.org/10.1016/j.csda.2017.03.003

[20] Sun, X.W., Song, X.Y. and Sun, L.Q. (2021) Additive Hazard Regression of Event History Studies with Intermittently Measured Covariates. *Canadian Journal of Statistics*, **50**, 511-531. https://doi.org/10.1002/cjs.11630

附　录

定理一证明

因为 $$\hat{\beta}-\beta_0=\hat{\beta}-\hat{A}^{-1}\hat{A}\beta_0=\frac{1}{n}\hat{A}^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{L}\left(Z_{il}-\bar{Z}_{kl}\right)\left[d_{ikl}-t_{ikl}\left(Z_{il}-\bar{Z}_{kl}\right)'\beta_0\right]$$
$$=\frac{1}{n}\hat{A}^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{L}\left(Z_{il}-\bar{Z}_{kl}\right)\mathrm{d}M_{ikl}$$。

对于任意有限的 k,l　$E\left(\mathrm{d}M_{ikl}\mid Z_{il}\right)=0$，$E\left(Z_{il}\mathrm{d}M_{ikl}\mid Z_{il}\right)=Z_{il}E\left(\mathrm{d}M_{ikl}\mid Z_{il}\right)=0$，$E\left(\left(Z_{il}-\bar{z}_{kl}\right)\mathrm{d}M_{ikl}\mid Z_{il}\right)=0$。由大数定理可知，$\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{L}\left(Z_{il}-\bar{Z}_{kl}\right)\mathrm{d}M_{ikl}\xrightarrow{P}0$，所以，由 $\hat{A}$ 的收敛性以及矩阵 A 的正定性，可以得到 $\hat{\beta}-\beta_0=\frac{1}{n}\hat{A}^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{L}\left(Z_{il}-\bar{Z}_{kl}\right)\mathrm{d}M_{ikl}\xrightarrow{p}0$，相合性得证。

下面证明渐近正态性，

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}-\beta_0\right)&=\hat{A}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{L}\left(Z_{il}-\bar{Z}_{kl}\right)\mathrm{d}M_{ikl}+o_p(1)\\&=\hat{A}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{L}\left(Z_{il}-\bar{z}_{kl}\right)\mathrm{d}M_{ikl}-\hat{A}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{L}\left(\bar{Z}_{kl}-\bar{z}_{kl}\right)\mathrm{d}M_{ikl}+o_p(1)\\&=\hat{A}^{-1}\cdot\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta_i+o_p(1)\end{aligned}$$

其中 $\eta_i=\sum_{k=1}^{K}\sum_{l=1}^{L}\left(Z_{il}-\bar{z}_{kl}\right)\mathrm{d}M_{ikl}$。这里的 η_i 是均值为 0，且独立同分布的随机变量，所以

$$Var\left(\sqrt{n}\left(\hat{\beta}-\beta_0\right)\right)\xrightarrow{p}A^{-1}Var\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta_i\right)A'^{-1}=A^{-1}Var\left(\eta_i\right)A'^{-1}=A^{-1}E\left(\eta_i^{\otimes 2}\right)A'^{-1}=A^{-1}\Omega A'^{-1}$$。

应用中心极限定理，可得 $n^{\frac{1}{2}}\left(\hat{\beta}-\beta_0\right)$ 的分布收敛到一个正态分布，这个正态分布的均值是 0，方差是 $A^{-1}\Omega A^{-1}$。$\Omega=E\left(\eta_i^{\otimes 2}\right)=E\left[\left(\sum_{k=1}^{K}\sum_{l=1}^{L}\left(Z_{il}-\bar{z}_{kl}\right)\mathrm{d}M_{ikl}\right)^{\otimes 2}\right]$

定理一证明

因为 $\mu_{0k}\left(t,\hat{\beta}\right)-\mu_{0k}(t)=\hat{\mu}_{0k}\left(t,\hat{\beta}\right)-\hat{\mu}_{0k}\left(t,\beta_0\right)+\hat{\mu}_{0k}\left(t,\beta_0\right)-\mu_{0k}(t)$，

$$\begin{aligned}\hat{\mu}_{0k}\left(t;\hat{\beta}\right)-\hat{\mu}_{0k}\left(t;\beta_0\right)&=\sum_{l=1}^{L}\left(\hat{\rho}_{kl}\left(\hat{\beta}\right)-\hat{\rho}_{kl}\left(\beta_0\right)\right)\left(\alpha_l\wedge t-\alpha_{l-1}\wedge t\right)\\&=\sum_{l=1}^{L}\left[\frac{\sum_{i=1}^{n}d_{ikl}-\sum_{i=1}^{n}t_{ikl}\hat{\beta}^{\mathrm{T}}Z_{il}}{\sum_{i=1}^{n}t_{ikl}}-\frac{\sum_{i=1}^{n}d_{ikl}-\sum_{i=1}^{n}t_{ikl}\hat{\beta}_0^{\mathrm{T}}Z_{il}}{\sum_{i=1}^{n}t_{ikl}}\right]\left(\alpha_l\wedge t-\alpha_{l-1}\wedge t\right)\\&=-\sum_{l=1}^{L}\left[\left(\hat{\beta}^{\mathrm{T}}-\hat{\beta}_0^{\mathrm{T}}\right)\frac{\sum_{i=1}^{n}t_{ikl}Z_{il}}{\sum_{i=1}^{n}t_{ikl}}\right]\left(\alpha_l\wedge t-\alpha_{l-1}\wedge t\right)\\&=-\sum_{l=1}^{L}\left(\hat{\beta}-\beta_0\right)'\bar{Z}_{kl}\left(\alpha_l\wedge t-\alpha_{l-1}\wedge t\right)\end{aligned}$$

$$\hat{\mu}_{0k}\left(t,\beta_0\right)-\mu_{0k}\left(t\right)=\sum_{l=1}^{L}\left[\frac{\sum_{i=1}^{n}d_{ikl}-\sum_{i=1}^{n}t_{ikl}\hat{\beta}^{\mathrm{T}}Z_{il}}{\sum_{i=1}^{n}t_{ikl}}-\rho_{0k}\left(t\right)\right]\left(\alpha_l\wedge t-\alpha_{l-1}\wedge t\right)$$

$$=\sum_{l=1}^{L}\left[\frac{\sum_{i=1}^{n}d_{ikl}-\sum_{i=1}^{n}t_{ikl}\hat{\beta}^{\mathrm{T}}Z_{il}-\rho_{0k}\left(t\right)\sum_{i=1}^{n}t_{ikl}}{\sum_{i=1}^{n}t_{ikl}}\right]\left(\alpha_l\wedge t-\alpha_{l-1}\wedge t\right)$$

$$=\sum_{l=1}^{L}\frac{\sum_{i=1}^{n}M_{ikl}}{\sum_{i=1}^{n}t_{ikl}}\left(\alpha_l\wedge t-\alpha_{l-1}\wedge t\right)$$

$$=\sum_{i=1}^{n}\frac{\sum_{l=1}^{L}M_{ikl}}{\sum_{i=1}^{n}t_{ikl}}\left(\alpha_l\wedge t-\alpha_{l-1}\wedge t\right)$$

由上一个定理的证明我们得到 $\hat{\beta}-\beta_0\xrightarrow{p}0$，所以 $\hat{\mu}_{0k}\left(t;\hat{\beta}\right)-\hat{\mu}_{0k}\left(t;\beta_0\right)\xrightarrow{p}0$，同样的 $\hat{\mu}_{0k}\left(t,\beta_0\right)-\mu_{0k}\left(t\right)=\dfrac{\frac{1}{n}\sum_{i=1}^{n}M_{ikl}}{\frac{1}{n}\sum_{i=1}^{n}t_{ikl}}\xrightarrow{p}0$ 所以得到 $\hat{\mu}_{0k}\left(t,\hat{\beta}\right)-\mu_{0k}\left(t\right)\xrightarrow{p}0$，所以相合性得证。

下面证明渐近正态性。

$$\sqrt{n}\left[\mu_k\left(t,\hat{\beta}\right)-\mu_{0k}\left(t\right)\right]=\sqrt{n}\left[\hat{\mu}_{0k}\left(t,\hat{\beta}\right)-\hat{\mu}_k\left(t,\beta_0\right)+\hat{\mu}_k\left(t,\beta_0\right)-\mu_{0k}\left(t\right)\right]$$

$$=\sum_{l=1}^{L}\left[-\sqrt{n}\left(\beta_0-\hat{\beta}\right)'\bar{Z}_{kl}+\sqrt{n}\frac{\sum_{i=1}^{n}\mathrm{d}M_{ikl}}{\sum_{i=1}^{n}t_{ikl}}\right]\left(\alpha_l\wedge t-\alpha_{l-1}\wedge t\right)$$

$$=\sum_{l=1}^{L}\left[-\frac{1}{\sqrt{n}}\cdot\left(\hat{A}^{-1}\sum_{i=1}^{n}\eta_i\right)'\bar{Z}_{kl}+o_P\left(1\right)+\sqrt{n}\frac{\sum_{i=1}^{n}\mathrm{d}M_{ikl}}{\sum_{i=1}^{n}t_{ikl}}\right]\left(\alpha_l\wedge t-\alpha_{l-1}\wedge t\right)+o_p\left(1\right)$$

交换连加顺序，并应用大数定律，上式可以转化为如下关系：

$$\sqrt{n}\left(\mu_{0k}\left(t,\hat{\beta}\right)-\mu_{0k}\left(t\right)\right)$$

$$=\sum_{i=1}^{n}\left[\left(-\frac{1}{\sqrt{n}}\left(\hat{A}^{-1}\sum_{l=1}^{L}\eta_i\right)'\bar{Z}_{kl}\right)\left(\alpha_l\wedge t-\alpha_{l-1}\wedge t\right)+\frac{1}{\sqrt{n}}\sum_{l=1}^{L}\frac{\mathrm{d}M_{ikl}}{\frac{1}{n}\sum_{i=1}^{n}t_{ikl}}\left(\alpha_l\wedge t-\alpha_{l-1}\wedge t\right)\right]+o_p\left(1\right)$$

$$=\sum_{i=1}^{n}\left[\left(-\frac{1}{\sqrt{n}}\sum_{l=1}^{L}\left(A^{-1}\eta_i\right)'\bar{z}_{kl}\right)\left(\alpha_l\wedge t-\alpha_{l-1}\wedge t\right)+\frac{1}{\sqrt{n}}\sum_{l=1}^{L}\frac{\mathrm{d}M_{ikl}}{\bar{t}_{kl}}\left(\alpha_l\wedge t-\alpha_{l-1}\wedge t\right)\right]+o_p\left(1\right)$$

$$=\sum_{i=1}^{n}\phi_{ik}+o_p\left(1\right)$$

这里，$\phi_{ik}(t)=\left(-\frac{1}{\sqrt{n}}\sum_{l=1}^{L}\left(A^{-1}\eta_i\right)'\bar{z}_{kl}\right)\left(\alpha_l\wedge t-\alpha_{l-1}\wedge t\right)+\frac{1}{\sqrt{n}}\sum_{l=1}^{L}\frac{\mathrm{d}M_{ikl}}{\bar{t}_{kl}}\left(\alpha_l\wedge t-\alpha_{l-1}\wedge t\right)$。

因为$E(\eta_i)=0$，所以$E\left(\sum_{l=1}^{L}\eta_i\right)=0$，则等式右边的第一项的期望为 0，同样的$E(\mathrm{d}M_{ikl})=0$，所以$E\left(\sum_{l=1}^{L}\mathrm{d}M_{ikl}\right)=0$，则等式右边第二项的期望为 0，所以有$E(\phi_{ik})=0$。又因为，对于固定的 t，上式$\sqrt{n}\left[\hat{\mu}_{0k}\left(t,\hat{\beta}\right)-\mu_{0k}(t)\right]$渐近于均值为零的独立同分布的随机变量之和，则由多元中心极限定理知，$\sqrt{n}\left[\hat{\mu}_{0k}\left(t,\hat{\beta}\right)-\mu_{0k}(t)\right]$以有限维分布弱收敛到均值为零的高斯过程，其在$(s,t)$处的协方差函数为$\Gamma_k(s,t)=\sum_{j=1}^{n}E\left(\phi_{jk}(s)\phi_{jk}(t)\right)$。