·基础理论与方法·

Cox及其拓展模型在基于队列的依时暴露因素效应估计中的应用

王振宇[1] 陈朔华[2] 赵欣宇[1] 王艳红[1] 吴寿岭[2] 王丽[1]

[1]中国医学科学院基础医学研究所/北京协和医学院基础学院流行病学和卫生统计学系，北京 100005；[2]开滦总医院，唐山 063000

通信作者：王丽，Email：liwang@ibms.pumc.edu.cn；吴寿岭，Email：drwusl@163.com

【摘要】 队列研究的特点之一是暴露因素会随时间而改变，如何充分利用暴露因素及其协变量的变化及其相互关系，从而获得更真实的暴露因素与结局关系是目前的研究热点。本研究以开滦队列为例，探讨基于基线暴露状态、随时间变化的暴露信息以及同时控制依时混杂因素时，如何利用Cox比例风险回归及其拓展模型，包括依时Cox回归及边际结构模型，探讨FPG与肝癌的关系，概述并比较了上述拓展模型的基本原理、应用条件、估计结果及结果解释。

【关键词】 队列研究；暴露变化；Cox比例风险回归；时间依赖性混杂；边际结构模型

基金项目：中国医学科学院医学与健康科技创新工程项目（2016-I2M-3-001）

DOI：10.3760/cma.j.cn112338-20200119-00046

Application of Cox and extended regression models on modeling the effect of time-updated exposures in cohort studies

Wang Zhenyu[1], Chen Shuohua[2], Zhao Xinyu[1], Wang Yanhong[1], Wu Shouling[2], Wang Li[1]

[1]Department of Epidemiology and Biostatistics, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences, School of Basic Medicine, Peking Union Medical College, Beijing 100005, China; [2]Kailuan General Hospital, Tangshan 063000, China

Corresponding authors: Wang Li, Email: liwang@ibms.pumc.edu.cn; Wu Shouling, Email: drwusl@163.com

【**Abstract**】 One of the characteristics on cohort studies is that exposures may change over time. The full use of information related to time-updated exposures, time-dependent covariates and their relationships to estimate the association between exposures and outcomes has become the hotspot of research. In this paper, the Kailuan cohort is used as an example to explore the association between fasting blood-glucose and hepatocellular carcinoma, based on different Cox regression models. Cox or time-dependent Cox regression models can be used to estimate the impact of exposure at baseline or on the time-updated exposures. When time-dependent confounders exist, marginal structure model is recommended. We also summarize the basic principles, conditions of applications, effect estimates, and results interpretation for each model, in this paper.

【**Key words**】 Cohort studies; Time-updated exposures; Cox proportional hazard model; Time-dependent confounding; Marginal structure model

Fund program: Innovation Fund for Medical Sciences, Chinese Academy of Medical Sciences (2016-I2M-3-001)

DOI:10.3760/cma.j.cn112338-20200119-00046

队列研究的主要特征之一是暴露因素及混杂因素均会随时间变化。在利用队列研究探讨暴露因素与结局关联关系时，常用方法之一是基于基线的暴露状态及协变量信息，在符合比例风险假说的前提下，采用Cox比例风险回归方法估计暴露与结局的风险关系[1]。但如何充分利用随访中暴露（time-updated exposures）和混杂因素的变化及其相互关系，从而获得更真实的暴露因素与结局之间的关联关系是当前研究热点。时间依赖Cox回归（time-dependent Cox regression model）常用于探讨暴露因素的变化与结局之间的关系[1]；但当同时考虑随时间变化的混杂因素对于结局的影响时，基于逆概率加权的边际结构模型（marginal structure models，MSMs）可能是更好的选择[2-3]。本研究将阐述Cox比例风险回归及其上述拓展方法的基本原理与应用条件，并基于开滦队列人群，系统阐述基于基

线FPG、FPG变化以及同时控制随时间变化的混杂因素时，如何利用上述方法探究FPG与肝癌的关系。

一、Cox回归及其拓展方法的基本原理

1. Cox比例风险回归基本原理：当研究目的为基于队列探讨基线暴露因素，如基线FPG与肝癌的发生关系时，Cox比例风险回归是最常用的方法，此时估计的是“暴露的长期效应”。该回归方法是将危险率函数与回归分析结合起来的一种方法，即：

$$\lambda(t,X,Z)=\lambda_0(t)exp(X\beta+Z\alpha) \quad (1)$$

其中t为观察对象发生结局事件时间的随机变量，X为暴露因素，Z为潜在混杂因素（图1A），$\lambda(t)$为危险率函数，即$\lambda(t)=\lim_{\Delta t\to 0}+\frac{pr(t\leqslant T<t+\Delta t\,|\,t\leqslant T)}{\Delta t}$，代表观察对象在$t$时刻结局事件的瞬间发生风险[4]。该模型需满足[4]：①比例风险假定，即$\frac{\lambda(t,X,Z)}{\lambda_0(t)}$不随时间的变化而变化；②模型中的协变量应与对数风险比呈线性关系，即对数线性假定。

基于上述模型，可以估计出暴露因素X与结局之间的关联效应风险比（Hazard ratio，HR）：

$$HR=\frac{\lambda(t,X_i,Z)}{\lambda(t,X_j,Z)} \quad (2)$$

其中X_i是暴露组取值，X_j是对照组取值。

2. 时间依赖Cox回归基本原理：利用基线暴露信息虽然能获得暴露对于结局发生的长期效应，但在真实的情况下，暴露因素可能会在随访过程中发生以下改变[1]：①基线的暴露水平与结局之间的效应关系会随时间而变化，此时可通过在常规的Cox回归模型中纳入依时协变量来估计暴露与结局之间效应；②暴露因素本身，如FPG会随着时间发生变化，而FPG的变化可能会影响未来结局如肝癌的发生（图1B），此时可通过时间依赖Cox回归，把整个随访时间划分为多个小区间，在每个小区间内利用开始时的暴露水平和结束时的结局进行Cox回归求出小区间的效应值，然后按照时间长度对每个区间的效应进行加权平均，最终获得整个随访期间内暴露水平对结局的效应大小（图2）。具体方法如下：

（1）数据拆分：首先依照原数据中的时间间隔t_m及暴露因素的测量次数m，将随访过程划分为m个时段（将图1B中的数据拆分为图2所示），每个时段为一条记录，共计m条。例如，第i个研究对象（$i=1,2,\cdots,n$），暴露因素X重复测量m_i次，则其随访数据被拆分为m_i次，其中第K时段内的记录包括第K时段的暴露信息X_{ki}、基线协变量Z_i和结局信息Y_{ki}（需要在区间结束时定义结局状态），以及该研究对象在第K时段的起始时间t_{0ki}和终止时间t_{1ki}（如果该个体在第K时段内发生结局，则该终止时间为其发生结局时间）。

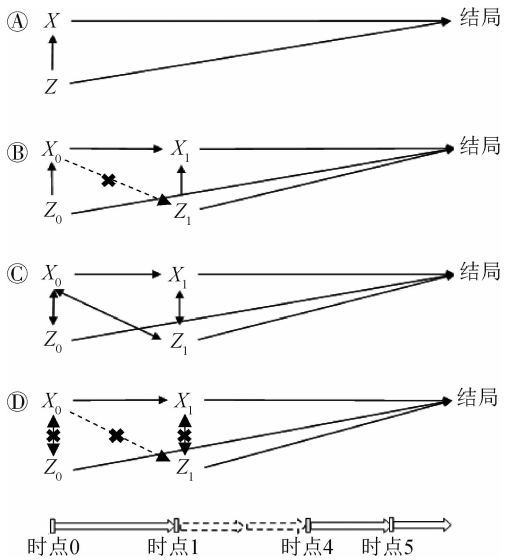

图1　基于队列的不同暴露因素与结局关联的病因

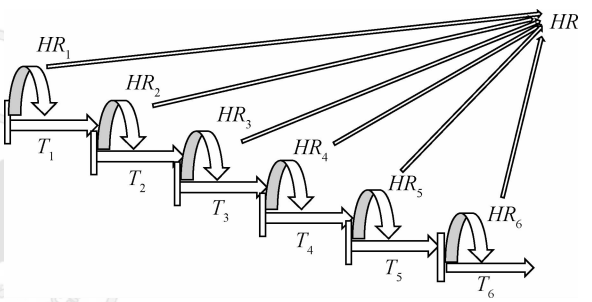

图2　基于时间依赖Cox回归的数据拆分及效应估计示意图

（2）构建Cox模型[5]：

$$\lambda(t,X_i,Z_i)=\lambda_0(t)*exp[X_i(t)^T\beta+Z_i^T\alpha]$$

$$=\lambda_0(t)*exp\left(\sum_{k=1}^{m_i}X_{ki}(t)^T\beta_k+Z_i^T\alpha\right) \quad (3)$$

其中，$\lambda_0(t)$为基准风险函数，假设除了离结局最近的记录t_i外，之前的重复测量记录与结局无关。

（3）采用偏极大似然函数估计暴露与协变量的参数[5]：

$$L(\alpha,\beta)=\prod_{i=1}^{D}\left\{\frac{exp[X_i(t_i)^T\beta+Z_i^T\alpha]}{\sum_{l\in R(t_i)}exp[X_l(t_i)^T\beta+Z_l^T\alpha]}\right\} \quad (4)$$

其中D为重复测量的时间点数，$t_1<t_2<t_3<\cdots<t_D$，其中$m_i\leqslant D$。

与传统的基于基线暴露信息估计“暴露对结局的长期效应”不同，时间依赖Cox回归估计的是“暴露对结局的短期效应”。该模型除了要求在每个小

区间内要满足Cox回归的2个假定外，在建模的过程当中，需要假设各区间之间的基线信息相互独立。

3. 边际结构模型基本原理：时间依赖Cox回归虽考虑了随时间变化的暴露因素对于结局的影响，但其假设前提是任意1个时间区间的协变量与其他时间区间内的协变量独立。实际研究中，如探讨血糖与肝癌的发生时，血糖与之后随访时点的其他协变量如ALT密切相关[6]，协变量不仅随时间而变化，同时又是暴露因素到结局发生的中间变量，即存在时间依赖性混杂（time-dependent confounder）[2,7]（图1C）。目前控制时依性混杂常用的方法为边际结构模型，该模型通过构建协变量及其历史信息与暴露因素的关系，估计每个个体进入不同暴露组的概率，并基于上述概率对样本人群中的每个个体进行逆概率加权（inverse probability weighting，IPW），最终建立虚拟总体，从而实现非条件下的因果推断。该方法在消除了其他协变量对暴露的影响同时，不改变暴露组和非暴露组的发病率，从而获得暴露和疾病之间的无偏估计（图1D）[2,8]。具体方法如下：

（1）数据拆分：与时间依赖Cox回归的拆分方法不同的是，需要依照确定的时间间隔t及暴露因素的测量次数m，按照研究对象的随访时长，将其拆分为从随访起点开始的m条不同随访时长的随访记录（如将图1C中的数据拆分为图3所示，记为$K=0,1,\cdots,m-1$）。例如，第i个研究对象（$i=1,2,\cdots,n$），其第K条记录包括其当前的暴露信息、协变量和结局信息为X_{ik}、Z_{ik}和Y_{ik}，随访时长信息$K\times t$，既往的暴露信息$\overline{X}_{(k-1)i}$，既往的协变量信息$\overline{Z}_{(k-1)i}$。

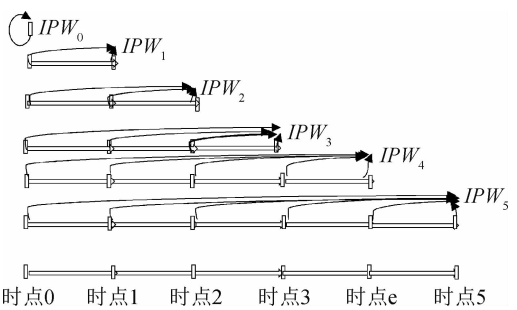

图3 基于边际结构模型的数据拆分及逆概率加权（IPW）估计

（2）逆概率权重估计：基于拆分后的数据库，通过拟合logistic回归估计每个个体在每个时点下的暴露状态的权重和失访权重。对于任意个体i的时点t的暴露因素X的稳定的逆概率加权的公式可表达为[2,7]：

$$IPW_i(t)=\prod_{k=0}^{int(t)}\frac{pr(X_k=x_{ki}\mid\overline{X}_{k-1}=\bar{x}_{(k-1)i},z_0=z)}{pr(X_k=x_{ki}\mid\overline{X}_{k-1}=\bar{x}_{(k-1)i},\overline{Z}_k=\bar{z}_{ki})}\quad(5)$$

其中，$\overline{X}_k=(X_0,X_1,\cdots,X_k)$为暴露的历史记录，同理$\overline{Z}_k$为其他协变量的历史记录，$\prod_{k=0}^{K}b_k=b_0\times b_1\times b_2\times b_3\times\cdots\times b_K$，$\overline{X}_{-1}=0$。

可通过logistic回归估计研究对象任意时点t进入暴露组的概率大小，此时模型需要考虑纳入所有与暴露因素相关的时依性混杂的历史记录$\overline{Z}_k$，以及暴露的历史信息$\overline{X}_{k-1}$：

$$\begin{aligned}&\text{logit}\,pr(X_k=1\mid\overline{X}_{k-1}=\bar{x}_{k-1},\overline{Z}_k=\bar{z}_{ki})\\&=\psi_0+\psi_1^T\bar{z}_k+\psi_2^T\bar{x}_{k-1}\end{aligned}\quad(6)$$

考虑到该概率的不稳定性，可以通过logistic回归拟合暴露所占的比例，此时模型估计时通常只需要纳入基线协变量信息Z_0和暴露的历史信息$\overline{X}_{k-1}$：

$$\begin{aligned}&\text{logit}\,pr(X_k=1\mid\overline{X}_{k-1}=\bar{x}_{k-1},Z_0=z)\\&=\varphi_0+\varphi_1^T z+\varphi_2^T\bar{x}_{k-1}\end{aligned}\quad(7)$$

长期随访过程，考虑到随访对象可能出现失访的情况，可进一步计算失访因素的逆概率权重（inverse probability censoring weight，IPCW）[2-3]。其计算方法同暴露因素的逆概率加权估计方法，此时结局变量为是否发生失访。最终每个个体每个时间点的稳定的逆概率权重（stable weight，SW）=IPW×IPCW。

（3）参数估计：在最终的总模型中，纳入随时间变化的暴露因素X_k与其他协变量的基线信息Z_0。可通过广义线性模型中的重复测量logit回归模型，赋予IPW和IPCW逆概率权重后，校正暴露因素及协变量信息后完成暴露因素的效应值估计。

边际结构模型的假设条件[3]：①可交换性：不存在未测量的混杂，即所有的混杂因素都需要被测量到。但是，这种假设很难进行测试，也无法评估没有测量到的变量效应，故通常认为该假设会是成立的；②正定性：在某种混杂因素存在的条件下，暴露因素出现的概率不能一直为0或1，如在职业流行病研究中，对在工作场所之外或不再工作的人来说，无法或不再接触到某种化学物质，暴露于该化学物质的概率为0，因此违反了正定性的假设；③正确估计权重函数。

二、实例分析

1. 案例信息：开滦队列是以开滦集团职工人群为基础的大型前瞻性研究队列[9]。于2006年建立后，每2年进行一次队列人群的健康体检（包括血生化检查、身体测量、B超检查）和问卷调查（包括一般人口学特征、个人疾病史、饮食生活习惯等），并通过开滦集团社会保障系统获得调查对象的肝癌发生结

局。每年约10万人完成调查，本研究基于2006—2017年已经完成的6次随访数据开展研究。

2. 基线FPG与肝癌发生风险：以2006—2007年的基线人群为研究对象，排除了基线无HBsAg和FPG信息者，既往或基线上发生肝硬化和肿瘤者，最终纳入分析95 418人，随访终点为肝癌发生、死亡或者到达研究终点（2017年12月31日）。

采用常规的Cox回归来探究高FPG与肝癌发生风险关系，其中多因素分析校正了基线时潜在的混杂因素（性别、年龄、糖尿病、饮酒、吸烟、体育锻炼、BMI、HBsAg、ALT、HDL-C、LDL-C、SBP、TG等）。结果显示，FPG≥5.6 mmol/L与FPG＜5.6 mmol/L相比，发生肝癌的风险增加33%，*HR*＝1.33（95%*CI*：1.01～1.76）（表1）。

3. 随时间变化的FPG与肝癌发生风险：以纳入对象中健康体检≥3次者（共77 496人）为研究对象，在不考虑时间依赖性混杂时，采用时间依赖Cox回归来探讨FPG与肝癌的关系。校正FPG和基线时潜在的混杂因素后，*HR*＝1.38（95%*CI*：0.99～1.94）。若存在时依性混杂，则采用边际结构模型，重新对数据进行拆分。结果显示，边际结构模型所得的权重*M*＝0.91（P_{25}～P_{75}：0.75～1.03）；经加权后，所有时间点的FPG≥5.6 mmol/L者与FPG＜5.6 mmol/L者相比，*HR*＝1.60（95%*CI*：1.13～2.26）。

三、讨论

在探讨随时间变化的暴露与结局的关系时，若关注的是暴露因素对结局的长期效应，可基于基线暴露信息采用常规Cox回归来解决，可以保证最大程度地利用队列人群，但不能应用随访过程中暴露因素及协变量的变化信息。而在没有时依性混杂的情况下，时间依赖Cox回归可以通过将整个随访时间划分为多个区间，每个区间基于该区间的暴露和结局信息进行效应值估计，从而达到充分利用暴露信息的变化来探讨暴露因素与结局的关联的目的；但由于各区间效应估计相对独立，只能反映暴露因素对结局的短期效应，同时未能充分考虑不同区间暴露信息和协变量的关联。

实际研究的大多数情况下，随时间变化的暴露与随时间变化的协变量密切相关。例如，在探讨FPG与肝癌关系时，不同区间的血糖水平以及影响血糖水平和肝癌发生的混杂因素如ALT等均密切相关而非完全独立，即存在时依性混杂。虽然时间依赖Cox回归可通过加权等方法，将多个时间依赖变量整合成一个新的时间依赖变量[10-11]，从而实现对多个随时间变化的变量进行综合分析，但仍不能真实的反映暴露的整个历史进程对于结局的影响，此时应选择边际结构模型。本研究中，多因素模型中基于基线FPG暴露的Cox回归和边际结构模型均观察到FPG与肝癌之间的显著关联，但在时间依赖Cox回归分析中，多因素结果无统计学意义。一方面可能因为短期效应假说对于FPG与肝癌风险而言不适用；另一方面，FPG与肝癌发生之间存在时依性混杂，提示应选择边际结构模型。而边际结构模型中校正时依性混杂后，*HR*＝1.60（95%*CI*：1.13～2.26），提示持续存在的FPG≥5.6 mmol/L者肝癌发生风险更高。

需要注意的是，在构建时间依赖Cox回归模型和边际结构模型中，均需要考虑[2-3]：①时间间隔的确定、每个个体在各区间的暴露信息，协变量信息以及结局信息的确定均取决于医学知识的判断和关键变量（包括暴露、结局信息及主要混杂因素）的重复测量频率。本研究实例中的所有数据均为2年更新一次，故时间间隔确定为2年，且每个区间的各变量的取值均为实际每次随访的测量值。但如果变量与变量的测量频率不同，此时需要先确定随访间隔，之后采用该时点前后某个点的数据或前后某一时间范围内的均值做为该区间的某变量的取值。②由于各区间的变量均不允许存在缺失，因此需要对数据进行填补，数据填补的准确性直接影响到效应的估计。除此之外，边际结构模型还需要注意准确获得稳定权重值是模型实施的前提和基础，过于极端的权重值会大幅影响最终模型结果，因此应保证权重值在1的附近波动，否则需采用权重截断的方法对极端权重进行校正[3,12-13]；此外，边际结构模型的使用必须满足正定性条件[3,12-13]，反之则需要使用其他的模型，如结构嵌套模型（structural nested models）[14-15]。

表1　基于不同Cox模型的FPG与肝癌关联效应估计

模　型	单因素分析			多因素分析		
	*HR*值(95%*CI*)	*Wald*χ^2值	*P*值	*HR*值(95%*CI*)	*Wald*χ^2值	*P*值
Cox回归	1.39(1.10～1.77)	7.317	0.007	1.33(1.01～1.76)	4.214	0.040
时间依赖Cox回归	1.65(1.19～2.28)	9.133	0.003	1.38(0.99～1.94)	3.496	0.062
边际结构模型	–	–	–	1.60(1.13～2.26)	7.051	0.008

注：参照组为FPG＜5.6 mmol/L

综上所述，在利用队列研究探讨暴露因素与结局关联关系时，当暴露信息随时间而变化时，我们需要基于不同的病因假说，以及不同时间点暴露因素与协变量以及结局的关系，选择合适的回归模型。如果关注的是基线暴露与未来结局发生的长期效应，此时推荐传统的Cox回归模型；如果考虑暴露的变化且更加关注暴露的短期效应时，可采用时间依赖Cox回归模型；若关注暴露的动态变化且期望控制时依性混杂时，推荐采用边际结构模型来探究暴露与结局之间的因果关联。

利益冲突 所有作者均声明不存在利益冲突

参 考 文 献

[1] Dekker FW, de Mutsert R, van Dijk PC, et al. Survival analysis: time-dependent effects and time-varying risk factors[J]. Kidney Int, 2008, 74(8): 994-997. DOI: 10.1038/ki.2008.328.

[2] Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology[J]. Epidemiology, 2000, 11(5): 550-560. DOI: 10.1097/00001648-200009000-00011.

[3] Xie DW, Yang W, Jepson C, et al. Statistical methods for modeling time-updated exposures in cohort studies of chronic kidney disease[J]. Clin J Am Soc Nephrol, 2017, 12(11): 1892-1899. DOI: 10.2215/CJN.00650117.

[4] Cox DR. Regression models and life-tables[J]. J Roy Stat Soc B Stat Methodol, 1972, 34(2): 187-202. DOI: 10.1111/j.2517-6161.1972.tb00899.x.

[5] Ngwa JS, Cabral HJ, Cheng DM, et al. A comparison of time dependent Cox regression, pooled logistic regression and cross sectional pooling with simulations and an application to the Framingham Heart Study[J]. BMC Med Res Methodol, 2016, 16(1): 148. DOI: 10.1186/s12874-016-0248-6.

[6] Hao C, Zhang C, Chen W, et al. Prevalence and risk factors of diabetes and impaired fasting glucose among university applicants in Eastern China: findings from a population-based study[J]. Diabet Med, 2014, 31(10): 1194-1198. DOI: 10.1111/dme.12473.

[7] Hernán Má, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men[J]. Epidemiology, 2000, 11(5): 561-570. DOI: 10.2307/3703998.

[8] 田丹平，张敏，黄渊秀，等. 边际结构模型基本原理及其应用实例介绍[J]. 中国卫生统计，2014，31(4)：725-728.
Tian DP, Zhang M, Huang YX, et al. Introduction to the basic principles of marginal structure model and its application[J]. Chin J Health Stat, 2014, 31(4): 725-728.

[9] Zhang Q, Zhou Y, Gao X, et al. Ideal cardiovascular health metrics and the risks of ischemic and intracerebral hemorrhagic stroke[J]. Stroke, 44(9): 2451-2456. DOI: 10.1161/strokeaha.113.678839.

[10] Fisher LD, Lin DY. Time-dependent covariates in the Cox proportional-hazards regression model[J]. Annu Rev Public Health, 1999, 20(1): 145-157. DOI: 10.1146/annurev.publhealth.20.1.145.

[11] Christensen E, Altman DG, Neuberger J, et al. Updating prognosis in primary biliary cirrhosis using a time-dependent Cox regression model[J]. Gastroenterology, 1993, 105(6): 1865-1876. DOI: 10.1016/0016-5085(93)91086-w.

[12] Williamson T, Ravani P. Marginal structural models in clinical research: when and how to use them?[J]. Nephrol Dial Transplant, 2017, 32 Suppl 2: ii84-90. DOI: 10.1093/ndt/gfw341.

[13] Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models[J]. Am J Epidemiol, 2008, 168(6): 656-664. DOI: 10.1093/aje/kwn164.

[14] Robins J, Tsiatis AA. Semiparametric estimation of an accelerated failure time model with time-dependent covariates[J]. Biometrika, 1992, 79(2): 311-319. DOI: 10.1093/biomet/79.2.311.

[15] Vansteelandt S, Joffe M. Structural nested models and G-estimation: the partially realized promise[J]. Stat Sci, 2014, 29(4): 707-731. DOI: 10.1214/14-STS493.

（收稿日期：2020-01-19）
（本文编辑：李银鸽）