arXiv:2603.02861v1 [stat.ME] 3 Mar 2026

半参数线性风险回归的聚焦信息准则

Axel Gandy（数学系，伦敦帝国理工学院）Nils Lid Hjort（数学系，奥斯陆大学）

2009年2月

摘要

McKeague 和 Sasieni（1994）提出的半参数线性风险回归模型，是对 Aalen（1980）所发展的线性风险回归模型的扩展。针对这类模型的模型选择方法仍不够成熟。在拟合半参数线性风险回归模型的过程中，通常从一组给定的协变量开始。对于每个协变量，至少有以下三种选择：允许其具有时变效应；允许其在时间上具有恒定效应；或将其从模型中排除。本文讨论聚焦信息准则（FIC），以帮助进行这种选择。按照 Claeskens和Hjort（2003）的理念，“聚焦”意味着人们关注某个特定的量，例如，具有某组特定协变量的患者在给定时间之前的生存概率。FIC需要估计所关注量的估计量的均方误差，并选择使该估计均方误差最小的模型。聚焦模型选择机制进一步扩展，以允许使用加权版本，从而形成适用的 wFIC方法，其目标是找出能够对给定参数列表提供良好估计的模型，例如某些患者子集的生存概率，或协变量向量某个指定区域内的生存概率。除了发展模型选择准则之外，本文还讨论了对最佳模型进行平均的相关方法。我们通过一个真实数据情境说明这些模型选择方法。

Key words and phrases: Aalen线性风险回归模型、聚焦模型选择、模型平均、半参数、生存分析

1 引言

我们考虑由Aalen（1980）提出的用于生存分析与事件史分析的线性加性风险回归模型，例如可参见Andersen等（1993，第8章）中的讨论以及Martinussen和Scheike（2006，第5章）。此处将 n 个体的风险率（或强度函数）建模为

$$h_i(s) = x_i^{\mathrm{t}}\alpha(s) = x_{i,1}\alpha_1(s) + \cdots + x_{i,q}\alpha_q(s) \quad \text{for } i = 1, \ldots, n, \tag{1}$$

就个体 i 的协变量信息 $x_i = (x_{i,1}, \ldots\ldots, x_{i,q})^{\mathrm{t}}$ 而言。通常，第一个协变量是截距，即 $x_{i,1} = 1$。回归函数 $\alpha_1(\cdot)\ldots\ldots, \alpha_q(\cdot)$ 是模型中的未知项，并假设其满足基本要求：对于所有 $s \in [0, \tau]$，以及协变量分布支持集中的所有向量 x，$x^{\mathrm{t}}\alpha(s)$ 均为非负数；此处 τ 是所涉及时间尺度的上限，通常为方便起见而选定。

Aalen 模型 (1) 最常以非参数方式应用，通常会得到累积回归函数的估计值和置信带

$$A_j(t) = \int_0^t \alpha_j(s)\,\mathrm{d}s \quad \text{for } j = 1, \ldots, q$$

这种方法避免对 $\alpha_j(\cdot)$函数施加参数限制；相关示例见第4节。不过，模型结构（1）也可以用于部分或全部回归函数取参数形式的情形。特别地，McKeague 和 Sasieni（1994）针对部分 $\alpha_j(\cdot)$采用非参数方法（例如针对 $j \in I$），而其余函数在时间上保持不变（例如针对 $j \in J$）的情况，提出了估计方法；其中， I和 J是 $\{1, \ldots, q\}$的互补子集。第2节将回顾所述估计策略。

关于Aalen 模型的理论和应用方面已有大量文献，但模型选择问题——例如应在分析中纳入哪些协变量——直到Hjort（2008）才似乎得到研究；该文提出了聚焦信息准则（FIC）。利用该FIC，可以从较长的候选协变量列表中决定应分别纳入或排除哪些协变量。对于给定的关注估计目标，该准则适用；例如，针对某类给定患者在某一时间点之后的生存概率，并通过估计每个可能考虑的 2^q 候选模型所涉及的均方误差来发挥作用。Hjort（2008）还提出了加权版本，其思想是通过适当的wFIC得分选择模型，从而对给定的估计目标列表产生精确估计，例如针对给定个体子集或协变量向量指定区域的生存概率，或针对给定个体在一系列未来时间点的生存概率。

本文的重点是将这些模型选择准则推广到每个协变量 x_j存在三种不同选项的情形：它可以非参数地纳入，即作为时变项；它可以参数地纳入，即随时间保持恒定；也可以被排除。因此，模型选择涉及确定三个互不相交的下标子集 $\{1, \ldots, q\}$，

– I，以时变效应纳入模型的那些；– J，以时间恒定效应纳入模型的那些；以及 – $K = \{1, \ldots, q\} - (I \cup J)$，从模型中排除的那些。

因此，总共有 3^q 个候选模型需要按质量排序。这样的模型选择器会得到如下形式的风险率

$$h_i(s) = \sum_{j \in I} x_{i,j}\alpha_j(s) + \sum_{j \in J} x_{i,j}\alpha_j \quad \text{for } i = 1, \ldots, n,$$

以及相应的累积危险率

$$H_i(s) = \sum_{j \in I} x_{i,j}A_j(s) + \sum_{j \in J} x_{i,j}\alpha_j t \quad \text{for } i = 1, \ldots, n, \tag{2}$$

以便与上文所述的半参数估计方法一起使用。本文所提出的模型选择准则侧重于估计累积危险率。

累积风险率的估计值常用于生成如下形式的生存概率估计值

$$\widehat{S}_i(t) = \prod_{[0,t]} \{1 - \mathrm{d}\widehat{H}_i(s)\},$$

无论是针对现有个体，还是针对具有相关协方差信息的某些预期患者类型，都是如此。因此，能够精确估计累积危险率的模型，也能精确估计生存概率。

FIC方法的一般思想见Claeskens和Hjort（2003），Claeskens和Hjort（2008a）对此进行了更广泛的讨论。针对生存数据的乘法Cox模型（其理念和实施方式均不同于本文的可加Aalen模型）所提出的FIC，以及加权的wFIC版本，由Hjort和Claeskens（2006）开发并加以阐释。

本文第2节提供了基本 FIC 得以发展的数学框架，并针对每个聚焦问题给出 3^q 候选模型的排序列表。特别是，不同的聚焦参数可能会产生不同的排序，并选出不同的最优模型。第3节发展了加权版本 wFIC，使得某些估计目标的重要性可以高于其他估计目标。随后，第4节对相关方法进行了说明，并将其应用于一个包含312名原发性胆汁性肝硬化（一种罕见肝病）患者的生存数据集。与其仅采用应用 FIC 或 wFIC 所建议的最优模型，不如对多个最优模型中的估计值进行加权平均。第5节简要讨论了模型平均这一主题。最后，第6节给出一些总结性评论，并提出进一步研究的问题。

2 理论发展

我们假设右删失生存数据的经典设置：我们观测到 n 个体。对于第 i个个体， T_i^0 表示事件时间， C_i 表示删失时间。我们只能观测到 T_i =min(T_i^0， C_i)和 $\delta_i = \mathrm{I}\{T_i \le C_i\}$，以及一个 q维协变量向量 x_i。观测事件所对应的0–1计数过程为 $N_i(t) = \mathrm{I}\{T_i \le t, \delta_i = 1\}$，其中 $i = 1, \ldots, n$。

加性模型 (1) 对应于计数过程 N_i，其强度函数为$\lambda_i(s) = R_i(s)x_i^{\mathrm{t}}\alpha(s)$，其中 $R_i(s) = \mathrm{I}\{T_i \ge s\}$ 是在险指示变量。特别地，随机过程 M_i 的增量为 $\mathrm{d}M_i(s) = \mathrm{d}N_i(s) - \lambda_i(s)\,\mathrm{d}s$，它们构成相互正交且均值为零的鞅，参见 Andersen 等人 (1993, 第 II.4 章)。这可简洁地表示为

$$\lambda(s) = Y(s)\alpha(s), \tag{3}$$

其中 $\lambda(s) = (\lambda_1(s), \ldots, \lambda_n(s))^{\mathrm{t}}$ 是 $N = (N_1, \ldots, N_n)^{\mathrm{t}}$的强度， $Y(s)$ 是一个 $n \times q$ 矩阵，其各行为 $Y_i(s) = R_i(s)x_i^{\mathrm{t}}$，而 $\alpha(s)$ 是模型中未知的 q 维参数函数。注意， $Y(s)$ 按鞅演算的术语来说是一个可料过程；其样本路径左连续，且在时间点 s 的取值在 s之前即可知。模型的参数空间是所有满足以下条件的 $\alpha(\cdot)$ 函数的集合：对于协变量向量分布的支撑集中的所有 x， $x^{\mathrm{t}}\alpha(s)$ 均为非负；特别是，根据该分布，某些 $\alpha_j(s)$ 分量可能在某些 s 取值区间上为负。

首先，我们希望估计协变量向量为 x 的个体的累积风险率，即

$$H(t, x) = x^{\mathrm{t}}A(t) = x_1A_1(t) + \cdots + x_qA_q(t),$$

其中， $A(t) = \int_0^t \alpha(s)\mathrm{d}s$ 是 $A_j(t) = \int_0^t \alpha_j(s)\,\mathrm{d}s$ 的向量。与其使用完整模型，我们不如使用部分协变量，和/或限制某些协变量具有恒定的时间效应，以降低估计的方差，正如引言中所解释的那样。更具体地说，我们希望采用 McKeague 和 Sasieni（1994）提出的半参数加性模型，其中

$$\lambda(s) = Y_I(s)\alpha_I(s) + Y_J(s)\alpha_J,$$

参见（2）。同样， I 是具有时变效应（非参数化建模）的指标集合， J 是具有时间不变效应（参数化建模）的协变量指标集合。

这里， Y_I（s）是 Y（s）的 $n \times |I|$ 子矩阵，由与集合 I中的协变量相对应的列组成（ $|I|$表示 I 的基数）； $n \times |J|$子矩阵 Y_J（s）的定义也类似。作为 H（t， x）的估计量，我们使用

$$\widehat{H}_{I,J}(t,x) = x_I^{\mathrm{t}} \widehat{A}_I(t) + x_J^{\mathrm{t}} \widehat{\alpha}_J t = \sum_{j \in I} x_j \widehat{A}_{I,j}(t) + \sum_{j \in J} x_j \widehat{\alpha}_{J,j} t, \tag{4}$$

^其中， $A_I(t)$ 和 $\widehat{\alpha}_J$ 是 McKeague 和 Sasieni（1994）提出的最小二乘估计量，即

$$\begin{aligned} \widehat{A}_I(t) &= \int_0^t \{Y_I(s)^{\mathrm{t}} Y_I(s)\}^{-1} Y_I(s)^{\mathrm{t}} \{\mathrm{d}N(s) - Y_J(s)\widehat{\alpha}_J \, \mathrm{d}s\}, \\ \widehat{\alpha}_J &= \Big\{ \int_0^\tau \widetilde{Y}_J(s)^{\mathrm{t}} \widetilde{Y}_J(s) \, \mathrm{d}s \Big\}^{-1} \int_0^\tau \widetilde{Y}_J(s)^{\mathrm{t}} \, \mathrm{d}N(s), \end{aligned}$$

w这里， $\widetilde{Y}_J(s) = [\mathrm{Id}_n - Y_I(s)\{Y_I(s)^{\mathrm{t}} Y_I(s)\}^{-1} Y_I(s)^{\mathrm{t}}] \, Y_J(s)$，其中 Id_n 为单位矩阵。

我们希望选择 I 和 J，使得估计量 (4) 的均方误差最小，即我们希望最小化

$$\begin{aligned} \mathrm{mse}_n(I,J) &= \mathrm{E}\{\widehat{H}_{I,J}(t,x) - H(t,x)\}^2 \\ &= \mathrm{Var}\ \widehat{H}_{I,J}(t,x) + \{\mathrm{bias}\, \widehat{H}_{I,J}(t,x)\}^2 \\ &= v_n(I,J) + \mathrm{sqb}_n(I,J), \end{aligned} \tag{5}$$

^记作 $\mathrm{sqb}_n(I, J)$，它是偏差 $\mathrm{E}\, H_{I,J}(t, x) - H(t, x)$ 的平方。我们的策略是根据数据估计这两项，最终选择 (I, J)，使相应的准则最小。

在接下来的推导中，我们假设加性模型 (3) 成立。首先将我们的估计量改写为

$$\widehat{H}_{I,J}(t,x) = \int_0^\tau K(s) \, \mathrm{d}N(s) = \sum_{i=1}^n \int_0^\tau K_i(s) \, \mathrm{d}N_i(s),$$

其中， $1 \times n$ 权重函数可以表示为

$$\begin{aligned} K(s) \;=\; & \mathtt{I}\{s \le t\} x_I^{\mathrm{t}} \{Y_I(s)^{\mathrm{t}} Y_I(s)\}^{-1} Y_I(s)^{\mathrm{t}} \\ & + \Big[t x_J^{\mathrm{t}} - x_I^{\mathrm{t}} \int_0^t \{Y_I(s)^{\mathrm{t}} Y_I(s)\}^{-1} Y_I(s)^{\mathrm{t}} Y_J(s) \, \mathrm{d}s \Big] \Big\{ \int_0^\tau \widetilde{Y}_J(s)^{\mathrm{t}} \widetilde{Y}_J(s) \, \mathrm{d}s \Big\}^{-1} \widetilde{Y}_J(s)^{\mathrm{t}} \end{aligned}$$

（在记号中省略其对 I、 J和 t 的依赖关系）。现在考虑如下分解

$$\widehat{H}_{I,J}(t,x) - H(t,x) = \Big\{ \int_0^\tau K(s) \, \mathrm{d}\Lambda(s) - H(t,x) \Big\} + \int_0^\tau K(s) \, \mathrm{d}\{N(s) - \Lambda(s)\}. \tag{6}$$

F对于第二项，可以利用鞅演算和中心极限定理证明

$$\int_0^t K(s)\{\mathrm{d}N(s) - \Lambda(s) \, \mathrm{d}s\} = \int_0^t K(s) \, \mathrm{d}M(s)$$

其分布接近一个均值为零的高斯过程，其方差函数 $V_n(t)$可估计为

$$\widehat{v}_n(I,J) = \int_0^\tau K(s) \, \mathrm{diag}\{\mathrm{d}N(s)\} \, K(s)^{\mathrm{t}} = \sum_{i=1}^n \mathtt{I}\{t_i \le t\} \delta_i K_i(t_i)^2.$$

更正式地说， $nV_n(t)$ 收敛到一个定义良好的极限方差函数，而 $n\widehat{v}_n(I, J)$ 是其相合估计量。

式 (6) 右侧第一项的期望值就是偏差。一种自然的偏差估计量为

$$\widehat{b}_n = \widehat{H}_{I,J}(t,x) - \widehat{H}_{\text{full}}(t,x),$$

其中

$$\widehat{H}_{\text{full}}(t,x) = \widehat{H}_{\{1,\ldots,q\},\emptyset}(t,x) = x^{\text{t}} \int_0^t \tilde{K}(s)\,\mathrm{d}N(s)$$

是累积风险的全模型估计量，其中 $\tilde{K}(s) = \mathrm{I}\{s \le t\}\{Y(s)^{\text{t}}Y(s)\}^{-1}Y(s)^{\text{t}}$是定义完整模型中 $H(t, x)$ 估计量时所涉及的权重函数。

因此，我们得到一个估计量 $\widehat{b}_n$，它是偏差 b_n 的无偏估计，其方差记为 τ_n^2，~可按上述方法估计，只需将 K 替换为 K。由此得到偏差平方的渐近无偏估计量：

$$\begin{aligned}\widehat{\text{sqb}}_n(I,J) &= \widehat{b}_n^2 - \widehat{\tau}_n^2 \\ &= \{\widehat{H}_{I,J}(t,x) - \widehat{H}_{\text{full}}(t,x)\}^2 - \int_0^\tau \{K(s) - \tilde{K}(s)\}\,\text{diag}\{\mathrm{d}N(s)\}\,\{K(s) - \tilde{K}(s)\}^{\text{t}}.\end{aligned}$$

回到式（5），由此得到聚焦信息准则

$$\text{fic}_n(I,J) = \widehat{v}_n(I,J) + \max\{0, \widehat{\text{sqb}}_n(I,J)\}, \tag{7}$$

如有需要，将偏差平方的负估计值截断为零。对于给定的协变量向量 x 和时间点 t，FIC方法是从所有候选模型中选择子集 I 和 J，使 $\text{fic}_n(I, J)$ 最小。

注释*A*。在各种应用中，坚持在最终模型中保留某些协变量是合理的，例如截距，也可能包括“性别”和“治疗或对照”。这相当于将某些协变量定义为“受保护”，从而限制潜在候选模型的列表。类似地，我们也可以坚持将某些协变量设为“受保护且随时间保持不变”，而将其他协变量设为“受保护且以非参数方式处理”。这不会影响本文所建立的方法，因为这仅意味着需要计算并排序的 FIC 值列表更短。

注释*B.* 上文中的FIC是针对从时间起点到某个给定 t的完整累积风险率估计而构建的。该机制与公式可轻易修改，以适用于形如 $H(t, x) - H(t_0, x)$的累积风险增量情形，这在研究已存活至时间 t_0的患者在时间 t 之后的生存概率时具有相关性。

注释*C.* 只要所需的矩阵逆存在，且公式产生非负且非递减的累积风险估计值，上述公式即有效。特别地，风险集内需要有足够多线性无关的协变量向量。该条件得到满足的概率通常接近1，并且只要在时间窗上限点 τ之上存在未删失生存时间的概率虽可能很小但为正，则随着样本量增大，该概率会以指数速度收敛到1。

3 加权FIC

上述FIC策略以（7）作为模型选择准则，但对于某些应用而言，可以说过于聚焦，因为它专门针对给定个体和给定时间点的情形。更一般地，人们可能希望选择一个适用于某一特定个体层（例如女性）且可能适用于一段时间范围（例如55至65岁年龄组）的模型。本节将提出一种加权FIC方法，用于在此类情况下选择合适的模型。

一般而言，考虑如下损失函数

$$\begin{aligned}\ell_n(I,J) &= \sum_k w_n(t_k,x_k)\{\widehat{H}_{I,J}(t_k,x_k) - H(t_k,x_k)\}^2 \\ &= \int \{\widehat{H}_{I,J}(t,x) - H(t,x)\}^2 \,\mathrm{d}W_n(t,x),\end{aligned}$$

在空间（t，x）中的各位置（t_k，x_k）上赋予特定权重 $w_n(t_k, x_k)$。积分记号表明，加权方案可以更为一般，例如在某个时间窗口内进行均匀加权等。我们希望将风险，即期望损失$r_n(I, J)$，降至最低。只要这些权重取决于协变量向量和时间点，而不以其他方式依赖数据，我们就有

$$r_n(I,J) = \mathrm{E}\ \ell_n(I,J) = \int \mathrm{mse}_n(I,J,t,x)\,\mathrm{d}W_n(t,x),$$

其中 $\mathrm{mse}_n(I, J, t, x)$ 如式（5）所示，但现在明确将（t，x）纳入记号。通过分别估计方差和偏差平方，可以自然地得到该量的估计值：

$$\widehat{r}_n(I,J) = \int \{\widehat{v}_n(I,J,t,x) + \widehat{\mathrm{sqb}}_n(I,J,t,x)\}\,\mathrm{d}W_n(t,x).$$

在这里，我们倾向于在必要时将累积的偏差平方估计值截断为零，而不是累积各个截断后的贡献。由此得到加权FIC选择准则

$$\mathrm{wfic}_n(I,J) = \int \widehat{v}_n(I,J,t,x)\,\mathrm{d}W_n(t,x) + \max\Big[0, \int \widehat{\mathrm{sqb}}_n(I,J,t,x)\,\mathrm{d}W_n(t,x)\Big].$$

备注 *D.* 以上考虑的权重函数假定为非随机函数；例如，它们可以取决于数据集的协变量向量，但不取决于随机寿命 t_i或其他估计参数。不过，在满足某些条件时，理论和公式也可以扩展到允许使用随机权重函数；参见 Claeskens和Hjort（2008b）。

注*E*。 一种特定的随机权重选择方式是，将 W_n 选为观测到的（t，x）值的经验分布（其中 t 限制在 $[0,\tau]$ 内）：

$$\mathrm{wfic}_n(I,J) = n^{-1}\sum_{i=1}^n \widehat{v}_n(I,J,\min(t_i,\tau),x_i) + \max\Big[0, n^{-1}\sum_{i=1}^n \widehat{\mathrm{sqb}}_n(I,J,\min(t_i,\tau),x_i)\Big].$$

这些权重的一个问题是，它们取决于删失程度。为避免这一点，我们建议，对于删失观测，实质上用完整模型在存活至时间 t_i这一条件下的条件分布来替代时间 t_i 。在实际操作中，这相当于从该条件分布中生成 r 个样本$t^*_{i,1}$，……，$t^*_{i,r}$，并在 $i = 1$，……，n和 $j = 1$，……，r上使用（$\min(t_{i,j}, \tau)\ x_i$）的经验分布。这种加权方式可视为一种很好的默认选择。

4 个示例

我们考虑Fleming和Harrington（1991）提出的著名PBC数据集，该数据集包含312名随机分组的原发性胆汁性肝硬化患者的随访数据。原发性胆汁性肝硬化是一种罕见的自身免疫性肝病。我们使用的协变量包括截距、治疗指示变量、水肿、性别、年龄、胆红素、白蛋白和凝血酶原时间。分类变量治疗指示变量、水肿是否存在以及性别（男性 $= 1$）均编码为0和1。协变量年龄、胆红素、白蛋白和凝血酶原时间均已进行中心化处理。

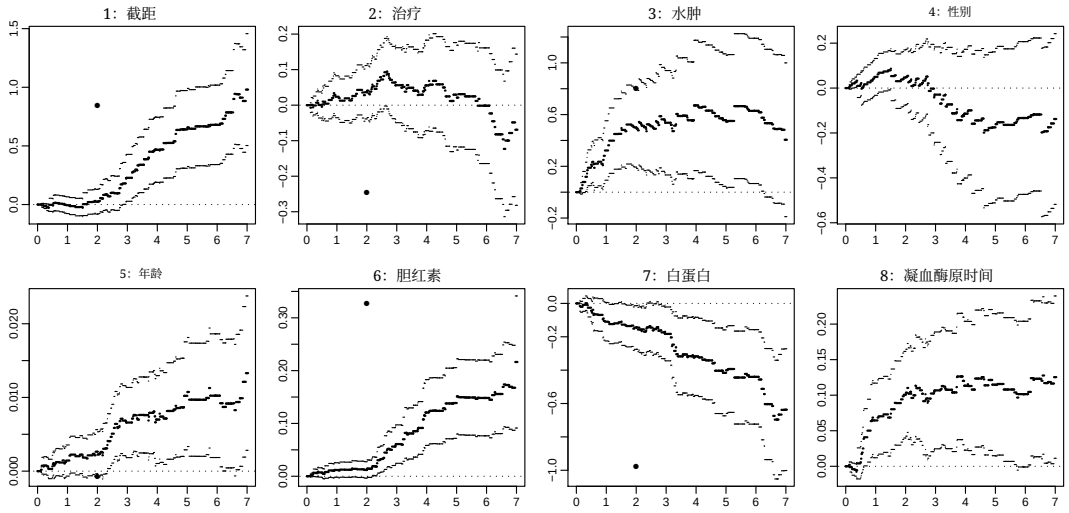

图1：PBC数据集：基于全加性模型的累积回归函数 $A_j(t)$的估计值，以及逐点95%渐近置信区间。时间以年为单位。

图1展示了完整模型的拟合结果。我们纳入了渐近逐点95%置信区间。在这些图以及后文中，我们使用端点 $\tau = 7$ 年。7年后，仍有30%的患者处于风险集中。

在下文中，我们将保护截距和治疗指示变量，即只考虑同时包含二者且二者均具有时变效应的模型。这样还剩下六个协变量，对于每一个协变量，我们都需要在允许时变效应、采用时间恒定效应以及将该协变量从模型中剔除之间做出选择。因此，我们考虑 $3^6 = 729$ 个潜在模型。

4.1 FIC

在本节中，我们考虑预测两个不同个体在时间 $t = 1$ 处的风险率 H 。

第一位个体是一名接受治疗的70岁男性，患有水肿。他的胆红素和凝血酶原时间在所有数据个体中均处于第75%分位数，而白蛋白处于第25%分位数。因此，他属于风险增加的个体（见图1）。

根据 FIC 对模型进行的排序以及 FIC图见图2。ˆ FIC图是针对所有考虑中的模型，绘制 FIC评分与预测 $H_{I,\ J}$ 之间关系的图。表格的第二列和第三列分别包含方差平方根和截断后的估计平方偏差 $\widehat{v}^{1/2}$ and $(\widehat{\ }\text{sqb}^{+})^{1/2} = \max(\widehat{\ }\text{sqb},\ 0)^{1/2}$，其中省略了对 n、 I 和 J 的依赖。当然，第二列和第三列的平方和恰好就是第1列中的 FIC，见式（7），

最优模型包含截距、治疗、性别和年龄这四个具有时间相依效应的协变量，并包含具有时间恒定效应的水肿。请注意，所有最优模型似乎都存在偏差。事实上，将最优模型的估计值与完整模型进行比较时，似乎$\widehat{\ }$最优模型给出的 H 估计值始终低于完整模型。

我们考虑的第二个个体同样是一名接受治疗的70岁男性，患有水肿，但其胆红素、白蛋白和凝血酶原时间均处于平均水平。根据FIC对模型进行排序的结果见表2。最优模型包含与之前相同的协变量，但现在假设所有未受保护的协变量都具有时间恒定效应。总体而言，模型的排序

fic	$\sqrt{\widehat{v}}$	$\sqrt{\widehat{\ }}$sqb + I		J	$\widehat{H}$
0.007	0.083	0.019	1,2,4,5	3	0.389
0.009	0.081	0.049	1,2,5	3	0.376
0.010	0.067	0.074	1,2,4	3,5	0.350
0.010	0.066	0.076	1,2	3,5	0.349
0.011	0.081	0.064	1,2,5	3,4	0.368
0.011	0.081	0.065	1,2,4,5	3,7	0.368
0.011	0.082	0.064	1,2,4,5,7	3	0.369
0.012	0.081	0.070	1,2,4,5	3,8	0.365
0.012	0.070	0.083	1,2,7	3,5	0.348
0.012	0.081	0.073	1,2,4,5,8	3	0.363
0.017	0.130	0.000	1,. . .,8		0.468

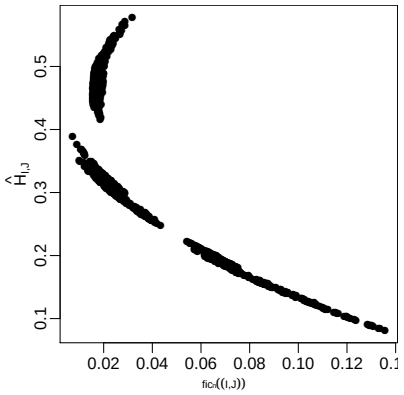

表1：PBC数据集：模型排名（最后一行为完整模型）及对应的FIC图。协变量编号与图1一致。我们考虑估计一名70岁男性在时间 $t = 1$ 、具有较高风险的胆红素、白蛋白和凝血酶原时间取值时的累积风险率。

fic	$\sqrt{\widehat{v}}$	$\sqrt{\widehat{\ }}$sqb + I		J	$\widehat{H}$
0.004	0.066	0.000	1,2	3,4,5	0.341
0.004	0.066	0.000	1,2	3,5	0.349
0.004	0.067	0.000	1,2,4	3,5	0.350
0.006	0.064	0.038	1,2,4	3	0.319
0.006	0.077	0.004	1,2,4,5,6	3	0.338
0.006	0.064	0.045	1,2	3	0.316
0.006	0.080	0.000	1,2,4,5	3,7	0.341
0.006	0.081	0.000	1,2,4,5	3,8	0.350
0.007	0.081	0.000	1,2,5	3,4	0.368
0.007	0.081	0.000	1,2,5	3	0.376
0.016	0.128	0.000	1,. . .,8		0.419

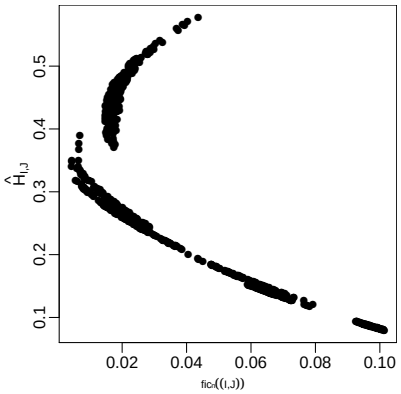

表2：设置与表1相同，但患者的胆红素、白蛋白和凝血酶原时间均为平均值。

与之前考虑的高风险个体相比有所不同 .

请注意，表2中的一些最优模型的估计平方偏差为0。事实上，在所有考虑的模型中，约有15%的模型的估计平方偏差为0。这并不完全令人意外，因为我们将估计平方偏差截断为0，而且我们只试图预测一个量（多个协变量被省略所导致的偏差可能会相互抵消）。

4.2 加权FIC

在本小节中，我们考虑将加权FIC（wfic$_n$）用作模型选择准则。权重 W_n 由 100 名“虚拟患者”的经验分布给出。这 100 名患者均为患有水肿的 50岁男性患者。我们将协变量向量的这部分称为 x_{fixed}。我们希望从条件分布 $x_{\text{rest}} \mid x_{\text{fixed}}$中选择其余协变量 x_{rest}。

我们假设当前的观测样本具有代表性，因此尤其是，针对这些 -

wfic	$\sqrt{\int \widehat{v}\mathrm{d}W}$	$\sqrt{(\int \widehat{\text{平方}}\,\mathrm{d}W)}$ + I		J	$\widehat{\text{准差}(H)}$	$\widehat{\text{均值}(H)}$ 标
0.015	0.099	0.073	1,2,4,8	3,5,6,7	0.512	0.138
0.015	0.098	0.076	1,2,8	3,5,6,7	0.510	0.138
0.017	0.101	0.081	1,2,4,6,8	3,5,7	0.505	0.120
0.017	0.117	0.057	1,2,6,7,8	3,5	0.533	0.126
0.017	0.117	0.058	1,2,4,6,7,8	3,5	0.533	0.127
0.017	0.101	0.085	1,2,6,8	3,5,7	0.502	0.119
0.018	0.097	0.092	1,2,8	3,4,5,6,7	0.499	0.138
0.018	0.117	0.067	1,2,7,8	3,5,6	0.532	0.145
0.018	0.117	0.069	1,2,4,7,8	3,5,6	0.532	0.144
0.019	0.116	0.073	1,2,6,7,8	3,4,5	0.524	0.126
0.029	0.170	0.000	1,. . .,8		0.634	0.104

表3：针对50岁男性水肿患者，在 $t = 1$ 时估计累积风险率时，模型按加权FIC排名。排名前10的模型。最后一行是完整模型。

wfic	$\sqrt{\int \widehat{v}\mathrm{d}W}$	$\sqrt{(\int \widehat{\text{平方}}\,\mathrm{d}W)}$ + I		J	$\widehat{\text{准差}(H)}$	$\widehat{\text{均值}(H)}$ 标
0.054	0.232	0.000	1,2	3,5,6,7	1.409	0.330
0.054	0.233	0.000	1,2	3,5,6,7,8	1.408	0.344
0.055	0.231	0.041	1,2,4	3,5,6,7,8	1.380	0.341
0.055	0.234	0.020	1,2,6	3,5,7	1.418	0.316
0.055	0.235	0.000	1,2,6	3,5,7,8	1.418	0.331
0.055	0.231	0.047	1,2	3,4,5,6,7,8	1.380	0.343
0.056	0.236	0.000	1,2,8	3,4,5,6,7	1.410	0.355
0.056	0.233	0.043	1,2,6	3,4,5,7,8	1.393	0.330
0.056	0.237	0.000	1,2,4,8	3,5,6,7	1.412	0.352
0.056	0.233	0.043	1,2,4,6	3,5,7,8	1.391	0.328
0.080	0.282	0.000	1,. . .,8		1.514	0.314

表4：在 $t = 3$ 时的累积风险率加权FIC分数。其他设置与表3相同。

个体的经验协方差矩阵能够合理反映总体的情况，并且协变量分布大致服从多元正态分布。基于这一假设，我们从适当的条件正态分布中为 100 名虚拟患者抽取 x_{rest} 。对于二分协变量（如治疗指示变量），正态性假设显然并不合理。为解决这一问题，我们将不受约束的实现值四舍五入到最接近的可能取值。

表3和表4分别展示了预测 $t = 1$ 和 $t = 3$ 年积分风险的结果。请注意，我们得到的“获胜”模型各不相同。此外，并没有明确的赢家；几个模型的 wfic_n 分数相当接近。

比较表3和表4中的最优模型，可以注意到，对于 $t = 1$ ，时变效应的协变量比对于 $t = 3$ 更多。事实上，对于 $t = 1$，所有最优模型都至少包含一个时变协变量，此外还有两个受保护协变量1和2。

$\widehat{\text{Mean}(H)}$ 列（相应地，SD(H) 列）包含该模型 100 个预测值的均值（相应地，标准差）。

图2展示了基于与表3相同 e

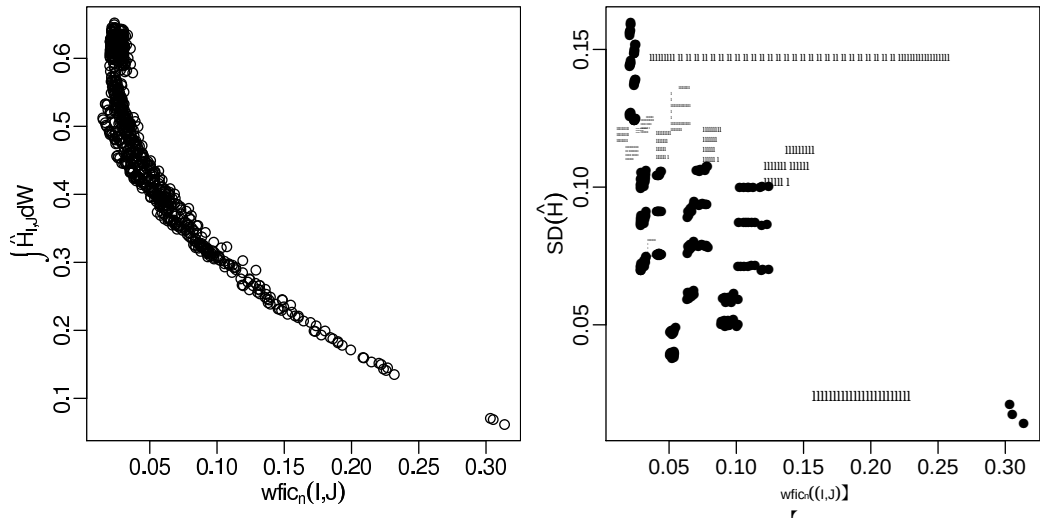

图2：对应表3的图。wFIC图及 wfic$_n$ 与所得估计值标准差的关系图。

的设置计算的加权FIC图的典型形式。图中绘出了wfic$_n$ 与平均预测值的关系。此外，还包含第二张图，展示了预测值的离散程度。这些图进一步表明，没有明确的胜出者，因此更凸显了模型平均技术的吸引力，我们将在第5节讨论这些技术。

4.3 经验加权FIC

表5列出了基于经验分布的模型FIC，具体方法见第3节末尾的注E，其中包含 $r = 1$。因此，我们使用权重（t_i^*，x_i）；对于未删失观测，$t_i^* = \min(t_i, \tau)$；对于删失观测，t_i^*是 τ与一个根据完整模型下给定 t_i的条件分布生成的随机变量中的较小值，并假设独立删失。

所有最优模型均纳入了协变量3（是否存在水肿）作为时变协变量。这与图1中完整模型的拟合结果一致：协变量3起初似乎有很强的影响，但大约在1.5年后这种影响便消失了。

5 模型平均

如上文给出的示例所示，通常有多个模型在FIC和wFIC尺度上的得分几乎相同。此时，人们很容易构造合适的模型平均估计量，例如采用以下形式：

$$\widetilde{H}(t,x) = \sum_{m=1}^{M} c(m)\widehat{H}_{I_m,J_m}(t,x),$$

其中，(I_m, J_m)（$m = 1, \ldots\ldots, M$）表示最有希望的 M个模型的指标集。权重 $c(m)$之和应为1，例如可以简单地取为 $1/M$，但自然应赋予得分较高的模型更大的权重。关于这类方法的一般讨论，参见 Hjort 和 Claeskens（2003）

wfic	$\sqrt{\int \widehat{v}\mathrm{d}W}$	$\sqrt{(\int \widehat{\quad}\mathrm{sqb}\ \mathrm{d}W)}$ + I		J	$\widehat{\text{准差}}(H)$	$\widehat{\text{均值}}(H)$ 标
0.024	0.138	0.066	1,2,3	5,6,7,8	0.390	0.460
0.025	0.140	0.071	1,2,3,5	6,7,8	0.388	0.459
0.025	0.142	0.068	1,2,3,7	5,6,8	0.394	0.467
0.025	0.139	0.075	1,2,3,8	5,6,7	0.394	0.451
0.026	0.146	0.066	1,2,3	4,5,6,7,8	0.391	0.464
0.026	0.144	0.071	1,2,3,5,7	6,8	0.392	0.466
0.026	0.142	0.077	1,2,3,7,8	5,6	0.397	0.459
0.026	0.141	0.079	1,2,3,5,8	6,7	0.391	0.450
0.027	0.146	0.073	1,2,3,8	4,5,6,7	0.394	0.456
0.027	0.147	0.072	1,2,3,4	5,6,7,8	0.392	0.461
0.031	0.176	0.000	1,. . .,8		0.402	0.517

表5：加权拟合。权重为所有个体的经验分布（针对删失进行了调整——见正文）。排名前10的模型。

此类模型平均估计量，其中加权方案的形式为

$$c(m) = \frac{\exp\{-\lambda\,\mathrm{fic}_n(I_m, J_m)\}}{\sum_{m'=1}^{M}\exp\{-\lambda\,\mathrm{fic}_n(I_{m'}, J_{m'})\}}$$

在最终风险函数方面，这些方法被推荐使用，并且已被证明效果良好。这里 λ 是一个算法参数，其值越高，就越重视最佳模型。由于所涉及的概率分布较为复杂，基于此类估计量的推断并不遵循标准方法，但可以从估计的最大模型中采用自助法；见第6节的注释A和B。

6 讨论和总结性评论

这最后一节提供一些补充性的总结性评论，其中一些可能会引发进一步的研究工作。

$\widehat{\ }$*A.* 最终估计量的行为。 文献中先前的研究已使每个基于子模型的估计量$H_{I,\ J}$（$t,\ x$）的性质得到相当充分的理解；特别是，当样本量较大时，每一个都近似服从正态分布，其方差可以一致地估计。当涉及最终选定的估计量时，问题则更为困难

$$\widetilde{H}(t,x) = \widehat{H}_{\widehat{I},\widehat{J}}(t,x),$$

因为所涉及的选定的子集也是随机的。其分布近似为相关正态分布的非线性混合（即多元正态向量的线性组合，其中权重依赖于该向量，因而本身也是随机的），特别是，它本身并不近似正态。因此，评估并估计其分布以及例如偏差和标准差，是相应具有挑战性的任务。第5节所考虑的更一般的模型平均估计量也伴随着类似的非平凡问题。可采用Hjort和Claeskens（203）以及Claeskens和Hjort（2008年a，第7章）的方法来处理这些问题。为了用检验推断陈述的置信区间来补充估计，可以使用自助法，我们将在下文加以说明。

$\widetilde{\ }$*B.* 自助法。 如上所述， $H(t,x)$ 的分布较为复杂。记

$$G_n(z, A_1, \ldots, A_q) = \mathrm{pr}\{\widetilde{H}(t,x) - H(t,x) \le z \mid A_1, \ldots, A_q\}$$

~误差 $H(t,x) - H(t,x)$ 的分布为，并在记号中体现其对未知回归函数 $A_1, \ldots, A_q$ 的依赖关系。我们现在建议通过代入完整模型中的相合估计量 $A_{1,\text{full}}, \ldots, A_{q,\text{full}}$ 来估计该分布，也就是说，使用分别对应于 $I = \{1, \ldots, q\}$ 和 $J = \emptyset$ 的非参数 Aalen 估计量。

这相当于从最大的估计模型进行自助法，结果为

$$\widehat{G}_n(z) = G_n(z, \widehat{A}_{1,\text{full}}, \ldots, \widehat{A}_{q,\text{full}}) = \text{pr}^*\{\widetilde{H}^*(t,x) - \widehat{H}_{\text{full}}(t,x) \le z \mid \widehat{A}_{1,\text{full}}, \ldots, \widehat{A}_{q,\text{full}}\},$$

~其中 pr* 表示数据（尤其是 H_{full} ）保持固定时所采用的概率机制 ~ ~ ^ ^，而 $H^*(t,x)$的计算方式与 $H(t,x)$完全相同，包括估计得到的子集 I^* 和 J^*，但使用的是形如(T_i^*, δ_i^*, x_i)的模拟数据集，其中 $i = 1, \ldots, n$。这里 $T_i^* = \min(T^{0,*}_i, C_i^*)$，其中 $T^{0,*}_i$ ^~从估计分布 $F_i(t) = 1 - \prod_{[0,t]}\{1 - \text{d}H_i(s)\}$中抽取，而 C_i^*要么是已知的删失时间（若适用），要么从 Kaplan–Meier 估计的删失分布中抽取。请注意，生成 T_i^*和 $\delta_i^* = \text{I}\{T_i^{0,*} \le C_i^*\}$时，$x_i$保持为其已知值。通过这种自助法，现在可以估计均方误差 $\text{E}\{\widehat{H}_{\widehat{I},\widehat{J}}(t,x) - H(t,x)\}^2$，并且^还可以设定置信区间，具体如下。令 $c_n < d_n$ 为上述 G_n~ ^ ~ ~中的 α 和 $1-\alpha$ 分位点，也就是 $H^* - H_{\text{full}}$的自助法分布中的分位点。那么，$[H - d_n, H - c_n]$ 是 H的近似$1 - 2\alpha$ 置信区间。这里假设最大的 Aalen 模型（即将模型中的每个 A_j 都视为时变函数的模型）是合适的。

C. 隐含检验。FIC和wFIC方法会生成按排名排列的候选模型列表，因此也隐含地回答了“是否应将协变量 x_j 从 J 移至 I？”这类问题。相应地，也隐含着一个检验统计量，用于检验 $\alpha_j(\cdot)$是否随时间保持恒定这一假设，并与其不恒定的备择假设相对。类似的问题还包括是否可能将 x_j 从 K 移至I，或从 K 移至 J；此时，模型选择方法可转化为检验假设 $\alpha_j = 0$ 的统计方法，并针对两种不同类型的备择假设（非零且恒定，以及非零且时变）进行检验。这些隐含检验确实具有由模型各方面决定的显著性水平，而且显著性水平未必很低——正如在参数模型中使用赤池信息准则AIC时，也隐含着相应的检验及其显著性水平。

D. 使用估计所得最优权重的估计量。本文所发展的方法基于McKeague 和 Sasieni（1994）的估计量，参见（4）。这些估计量可视为“直接基于鞅的最小二乘估计量”。不过，也可以采用更复杂的加权的基于鞅的最小二乘估计量，其中的权重是理论最优权重的估计值；而要得到这些权重，必然还需要对（1）中的回归函数 α_1、.....、α_q 本身进行非参数估计（而非估计其累积函数）。我们的模型选择方法也可以扩展到涵盖这些更精细的估计量。不过，这套方法在技术上要求高得多，而且即便可以证明在样本量非常大时能确保精度提高，精度究竟能提高多少也并不明确，参见Huffer和McKeague（1991）。

*E. FIC*相对准则。我们通过将平方偏差的基础估计截断为零，构造了 FIC 准则 (7)。不过，也可以保留平方偏差估计而不作任何改动，由此得到略有不同的 FIC相对准则，例如 $\text{fic}_n^*(I, J)$，而不是 $\text{fic}_n(I, J)$。这两种方案（以及其他方案）孰优孰劣，可以从不同角度加以研究——例如，可以考察选中基础最优子模型的概率，^或最终估计量 $H_{\widehat{I},\widehat{J}}$的精度。Claeskens和Hjort（2003）进行了一些此类分析和比较，结果表明，很难断定哪种方案整体上占优；两者各自得到的风险函数只是在参数空间的不同区域占优。

致谢

作者谨向邀请他们参与奥斯陆科学与文学学院高级研究中心这一充满活力的学术环境的各位表示感谢。Odd Aalen 和 Ørnulf Borgan 在此为 2005–2006 学年组织了“复杂事件历史数据的统计分析”项目。

Aalen, O.O. (1980). 一种用于计数过程非参数回归分析的模型。载于数学统计与概率论——第六届国际会议论文集，维斯瓦（波兰），统计学讲义第2卷，第1–25页。施普林格出版社，纽约。Andersen, P., Borgan, Ø., Gill, R. 和 Keiding, N. (1993). 基于计数过程的统计模型。施普林格出版社，纽约。Claeskens, G. 和 Hjort, N.L. (2003). 聚焦信息准则。美国统计学会会刊 **98**，901–916。Claeskens, G. 和 Hjort, N.L. (2008a). 模型选择与模型平均。 剑桥大学出版社，剑桥。Claeskens, G. 和 Hjort, N.L. (2008b). 回归模型中的平均风险最小化。计量经济学理论 **24**，493–527。Fleming, T. 和 Harrington, D. (1991). 计数过程与生存分析。威利出版社，纽约。Hjort, N.L. 和 Claeskens, G. (2003). 频率学派模型平均估计量。美国统计学会会刊**98**，879–900。Hjort, N.L. 和 Claeskens, G. (2006). Cox风险回归模型的聚焦信息准则与模型平均。美国统计学会会刊 **101**，1449–1464。Hjort, N.L. (2008). 线性风险回归模型的聚焦信息准则。载于生物医学与技术系统的统计模型与方法，第487–502页。Birkh¨auser，波士顿。Huffer, F.W. 和 McKeague, I.W. (1991). Aalen加性风险模型的加权最小二乘估计。美国统计学会会刊 **86**，114–129。Martinussen, T. 和 Scheike, T. (2006). 生存数据的动态回归模型。施普林格出版社，海德堡。McKeague, I. W. 和 Sasieni, P. D. (1994). 一种部分参数加性风险模型。生物统计学**81**，501–514。