arXiv:2601.03675v1 [stat.ME] 7 Jan 2026

用于组合多项诊断测试的最大平滑似然方法及其在受试者工作特征估计中的应用

Fangyong Zheng[1], 、Pengfei Li[2],和 Tao Yu[3]

[1]Fangyong Zheng 是新加坡国立大学统计与数据科学系博士生，新加坡 117546（电子邮箱：e1117764@u.nus.edu）[2]Pengfei Li 是滑铁卢大学统计与精算科学系教授，加拿大安大略省滑铁卢市 N2L 3G1（电子邮箱：pengfei.li@uwaterloo.ca）[3]Tao Yu 是新加坡国立大学统计与数据科学系副教授，新加坡 117546（电子邮箱：yu.tao@nus.edu.sg）

摘要

在医学诊断中，与使用单一生物标志物相比，利用多种生物标志物可以显著提高分类准确率。尽管现有基于指数倾斜或密度比模型的方法已显示出良好前景，但其假设在实践中可能过于严格。本文采用一种灵活的半参数模型，通过生物标志物线性组合的未知单调变换来描述患病受试者与健康受试者之间的密度比。为提高估计效率，我们提出一个利用基础密度及变换函数平滑性的平滑似然框架。基于最大平滑似然方法，我们构造模型参数及相应概率密度函数的估计量。我们开发了一种有效的计算算法以便实施，推导所提出估计量的渐近性质，并建立受试者工作特征（ROC）曲线及曲线下面积（AUC）的估计方法。通过模拟研究和真实数据应用，我们证明，与现有方法相比，所提出的方法能够得到更准确、更有效的估计值。

关键词：生物标志物的最优组合；受试者工作特征；平滑似然；形状约束推断；加权保序回归

1 引言

在医学诊断中，通常会为每位受试者收集多个生物标志物。每种生物标志物都可能携带有关疾病是否存在、严重程度或进展情况的部分且互补的信息。依赖单一生物标志物可能导致评估不完整或不准确，尤其是在疾病具有异质性或影响多个生理通路的情况下。因此，适当纳入并组合多个生物标志物，对于提高诊断准确性、优化风险分层以及实现更加个性化的医疗决策，变得日益重要（Kodosaki 等，2024年；Kim 等，2021年）。有效的组合策略可以利用生物标志物的联合分布，构建强大的诊断工具，例如综合评分或多变量ROC曲线，从而更好地区分患病个体与未患病个体；参见Ghosal （2025年）及其中的参考文献。

提出适当的组合方法、建立有效的数值算法并探究其理论性质，数十年来一直是统计学界的热门研究课题。假设生物标志物服从多元正态分布，Suand Liu (1993) 建立了最佳线性组合方法。基于 Neyman–Pearson 引理，McIntoshand Pepe (2002) 指出，患病受试者与健康受试者之间的密度比，在任意固定假阳性率下都能使真阳性率达到最大，因此成为组合多种生物标志物的金标准；另见 Eguchiand Copas (2002);Copas andCorbett (2002)。文献中已有基于密度比模型的生物标志物组合方法。 Qinand Zhang (2010) 考虑了一种指数倾斜模型，其中假设患病与未患病人群之间的对数密度比是多项诊断测试的线性组合；具体而言，该模型表示为

人群是多项诊断测试的线性组合；具体而言，该模型表示为

$$\frac{f(x)}{g(x)} = \exp\left\{\alpha + \beta^{\tau}\xi(x)\right\}, \tag{1}$$

其中 α 是一个标量参数， $x \in \mathbb{R}^d$， β 是一个 p 维参数， $\xi(\cdot)$ 是一个已知的 $p \times 1$关于 x 的光滑向量函数， $f(\cdot)$ 和 $g(\cdot)$ 分别表示患病受试者和健康受试者的生物标志物测量值的联合概率密度函数（概率密度函数）。Chen等人(2016年)指出，(1)所假设的指数函数形式可能过于严格，在实践中未必能很好地满足；示例见第6 节。他们提出了以下半参数模型:

$$\frac{f(x)}{g(x)} = \phi(\beta^{\tau}\xi(x)), \tag{2}$$

其中 $\phi(\cdot)$ 是一个未知的非递减函数，并提出了一种基于经验似然的方法来估计模型参数。关于组合生物标志物的方法的更多进展，参见Ma和Huang (2007), Barreno等人 (2008年), Liu 和 Zhou (2013 年), Kim 等人 (2013 年), Huang 和 Sanda (2022 年等。

模型（2）的限制性弱于指数倾斜模型（1）；当模型（1）所依据的假设不成立时，模型（2）可以对 β 给出更准确的估计。因此，本文采用模型（2）。Chen 等（2016）提出的经验似然方法虽然适用于该模型，却未能利用潜在密度 f（x）和 g（x）或单调变换函数 ϕ（·）的潜在平滑性。然而，在许多实际应用中，这些函数确实是平滑的，未能利用这一性质可能会导致效率损失和估计结果欠佳。为解决这一问题，我们提出一种基于模型（2）并纳入平滑性的估计方法，以提高对 f（·）、 g（·）和β的估计精度。具体而言，我们采用平滑似然方法，该方法基于最大平滑

似然方法，由 Eggermont 和 LaRiccia (1995a) 提出。这一技术与参数设定下的最大似然估计相类似，但专为非参数密度估计设计，同时保留了一致性和效率等诸多理想性质。平滑似然方法已被证明在各种具有挑战性的统计问题中行之有效，包括反卷积 Eggermont 和 LaRiccia (1995b）、平滑单调及单峰密度估计 Eggermont 和 LaRiccia (2000）、风险率和事件时间分布估计 Groeneboom 等 (2010）、混合模型中的密度估计 Levine 等人 (2011);Yu 等 (2019）、区间删失数据 Groeneboom (2014），以及具有似然比序的双样本问题 Yu 等 (2017）。我们基于这些进展，将平滑似然框架纳入估计程序。

本文假设模型（2），提出平滑似然，并基于此定义$\widehat{\quad}$ $\widehat{\quad}$估计量 β， $\widehat{g}$（·）和 f（·），分别用于参数 β， g（·）和 f（·）。我们还建立并实现了一种高效的计算算法，用于估计这些参数。利用这些估计量，我们进一步估计ROC曲线和曲线下面积。我们严格推导了参数估计量以及ROC和曲线下面积估计量的渐近性质。通过大量模拟研究，我们证明，与其他竞争方法相比，我们的方法具有更高的估计准确性。最后，我们通过一个真实数据实例展示了我们的方法的实用价值。

本文其余部分的组织如下。第2 节介绍所提出的最大平滑似然方法，并定义相应的参数估计量。第3节介绍参数估计的计算算法。第4 节建立所提出估计量的渐近性质。第5 节描述ROC曲线和曲线下面积的估计，并研究其渐近行为。第6 节报告模拟研究结果，第7 节使用胰腺癌数据集说明所提出的方法。第8

节通过讨论对本文进行总结。技术条件见附录，而支持第2、4和5 节理论结果的详细技术推导则收录于补充材料中。

2 最大平滑似然估计

设 $\{X_1, \ldots, X_n\}$ 和 $\{Y_1, \ldots, Y_m\}$ 分别为来自患病者和健康受试者的 d维检测结果。假设 $X_1, \ldots, X_n$ 独立同分布，服从随机向量 X的分布； $Y_1, \ldots, Y_m$ 独立同分布，服从 Y的分布。用 $F(x)$ 和 $f(x)$分别表示 X的累积分布函数和概率密度函数；同理，用 $G(x)$ 和 $g(x)$ 分别表示 Y的累积分布函数和概率密度函数。我们采用方程（2）中的单调密度比模型。为便于表述，且不失一般性，我们考虑以下形式：

$$\frac{f(x)}{g(x)} = \phi(\beta^\tau x), \tag{3}$$

其中， $\phi(\cdot)$是定义在 $\mathbb{R}$上的未知非递减函数，因此 $p = d$。为确保模型的可识别性，我们假设 β 的第一个分量等于 1。忽略常数项后的对数似然为

$$\widetilde{l}_{n,m}(f, g) = \frac{1}{n+m}\left[\sum_{i=1}^{n} \log\{f(X_i)\} + \sum_{j=1}^{m} \log\{g(Y_j)\}\right], \tag{4}$$

满足（3），并且

$$f(x) \geq 0, \quad g(x) \geq 0, \quad \int_{\mathbb{R}^d} f(x)dx = \int_{\mathbb{R}^d} g(x)dx = 1. \tag{5}$$

然而，该对数似然是无界的，因为我们可以使 f（X_i）对于每个 X_i任意增大。可以借用平滑似然方法来处理这一无界问题；参见Yu 等人（2017、2019）；Levine 等人（2011）及其中的参考文献。具体而言，令 $N_H f$（x）为作用于密度 f（·）的非线性平滑算子，其定义为

$$N_H f(x) = \exp\left\{\int_{\mathbb{R}^d} K_H(u-x)\,\log f(u)\,du\right\},$$

其中 $K_H(x) = |H|^{-1}K(H^{-1}x)$。这里，$K(\cdot)$ 是对称多元概率密度函数，用作核函数；H 是 $d \times d$ 对角正定带宽矩阵。我们分别将对数似然（4）中的 f 和 g 替换为$N_{H_1} f$ 和 $N_{H_2} g$，得到平滑对数似然：

$$l_{n,m}(f,g) = \frac{1}{n+m}\left[\sum_{i=1}^{n}\log\{N_{H_1}f(X_i)\} + \sum_{j=1}^{m}\log\{N_{H_2}g(Y_j)\}\right], \tag{6}$$

其中 H_1 和 H_2 是带宽矩阵，H_1、$H_2 \to 0$ 作为 $n \to \infty$。

~令 $f_n(x) = n^{-1}\sum_{i=1}^{n} K_{H_1}(x - X_i)$、$\widetilde{g}_m(x) = m^{-1}\sum_{j=1}^{m} K_{H_2}(x - Y_j)$，以及 $\lambda = n/(n+m)$；为简便起见，我们假设 $\lambda \in (0,1)$ 且为常数。式（6）中的平滑对数似然 l_n、$_m(f, g)$ 可重写为

$$l_{n,m}(f,g) = \lambda\int_{\mathbb{R}^d}\widetilde{f}_n(x)\log\{f(x)\}\,dx + (1-\lambda)\int_{\mathbb{R}^d}\widetilde{g}_m(x)\log\{g(x)\}\,dx. \tag{7}$$

经过重新参数化：

$$\theta(\beta^\tau x) = \frac{\lambda f(x)}{\lambda f(x) + (1-\lambda)g(x)} = \frac{\lambda\phi(\beta^\tau x)}{\lambda\phi(\beta^\tau x) + (1-\lambda)}, \quad \psi(x) = \lambda f(x) + (1-\lambda)g(x),$$

得到

$$f(x) = \frac{\theta(\beta^\tau x)\psi(x)}{\lambda}, \qquad g(x) = \frac{\{1-\theta(\beta^\tau x)\}\,\psi(x)}{1-\lambda}, \tag{8}$$

我们可以将（7）中的平滑对数似然函数写为

$$l_{n,m}(f,g) = l_{n,m,1}(\theta,\beta) + l_{n,m,2}(\psi) + \text{constant},$$

其中

$$\begin{aligned} l_{n,m,1}(\theta,\beta) &= \int_{\mathbb{R}^d} \lambda\widetilde{f}_n(x)\log\theta(\beta^\tau x)dx + \int_{\mathbb{R}^d}(1-\lambda)\widetilde{g}_m(x)\log\{1-\theta(\beta^\tau x)\}dx \\ l_{n,m,2}(\psi) &= \int_{\mathbb{R}^d}\left\{\lambda\widetilde{f}_n(x) + (1-\lambda)\widetilde{g}_m(x)\right\}\log\psi(x)dx. \end{aligned} \tag{9}$$

下文将把 $\theta(\cdot)$ β、 $\psi(\cdot)$ 作为关注参数；基于（3）和（5），我们的方法针对 $\theta(\cdot)$ β、 $\psi(\cdot)$ 提出的估计量，是 $l_{n、m、1}(\theta,\beta)$ 和 $l_{n、m、2}(\psi)$ 的极大化者，满足

$$\begin{aligned} &\psi(\cdot) \geq 0, \quad \int_{\mathbb{R}^d}\psi(x)dx = 1; \quad \int_{\mathbb{R}^d}\psi(x)\theta(\beta^\tau x)dx = 1\,; \\ &\theta(\cdot) \in [0,1] \text{ and is non-decreasing.} \end{aligned} \tag{10}$$

$\psi(\cdot)$ 和 $(\theta(\cdot)\ \beta)$ 的估计可以分别基于 $l_{n,\ m,\ 2}(\psi)$ 和 $l_{n,\ m,\ 1}(\theta,\beta)$ 进行。参见 Eggermont 和 Lariccia (2001年) 第 122 页，$l_{n,\ m,\ 2}(\psi)$ 的最大化解为

$$\widehat{\psi}_{n,m}(x) = \lambda\widetilde{f}_n(x) + (1-\lambda)\widetilde{g}_m(x), \tag{11}$$

它作为 $\psi(\cdot)$ 的估计量。 $\theta(\cdot)$ β 的估计量定义为

$$(\widehat{\theta}(\cdot), \widehat{\beta}) = \arg \max_{\theta \in \Theta, \beta \in \mathbf{B}} l_{n,m,1}(\theta, \beta), \tag{12}$$

其中 $\Theta = \{\theta(\cdot) : \theta(\cdot) \in [0,1]$ 且为非递减函数$\}$ 且 $\mathbf{B} = \{1\} \times \mathbf{B}_{-1}$ 分别是 $\theta(\cdot)$ 和 β的参数空间。优化问题 (12) 可通过以下剖面化程序求解。对于每个 $\beta \in \mathbf{B}$，令

$$\theta_\beta(\cdot) = \arg \max_{\theta \in \Theta} l_{n,m,1}(\theta, \beta). \tag{13}$$

下面的定理 1 保证优化问题 (13) 等价于连续加权保序回归。

定理1。 假设满足附录中的条件 *1* 。对于每个 $\beta \in \boldsymbol{B}$，考虑 $\theta_\beta(\cdot)$ ，其定义见 (13)。

(a) 等价地， $\theta_\beta(\cdot)$ 也是连续加权保序回归的解：

$$\theta_\beta(\cdot) = \arg \min_{\theta \in \Theta} \int_{\mathbb{R}^d} \left\{ \frac{\lambda \widetilde{f}_n(x)}{\widehat{\psi}_{n,m}(x)} - \theta(\beta^\tau x) \right\}^2 \widehat{\psi}_{n,m}(x) dx. \tag{14}$$

(b) 考虑 $\widehat{\psi}_{n,\ m}(\cdot)$由 (11)给出。我们有

$$\int_{\mathbb{R}^d} \widehat{\psi}_{n,m}(x) \theta_\beta(\beta^\tau x) dx = \lambda.$$

定理1 的（a）部分表明， $\theta_\beta(\cdot)$ 是连续加权保序回归的解（Groeneboom和 Jongbloed， 2010），可以很容易地通过数值方法求解；具体细节

见第3节。随后，剖面似然表示为

$$l^*(\beta) = \int_{\mathbb{R}^d} \lambda \widetilde{f}_n(x) \log \theta_\beta(\beta^\tau x) dx + \int_{\mathbb{R}^d} (1-\lambda) \widetilde{g}_m(x) \log \{1 - \theta_\beta(\beta^\tau x)\} dx, \tag{15}$$

因此

$$\widehat{\beta} = \arg\max_{\beta \in \mathbf{B}} l^*(\beta) \quad \text{and} \quad \widehat{\theta}(\cdot) = \theta_{\widehat{\beta}}(\cdot).$$

$\widehat{\ }$ $\widehat{\ }$ $\widehat{\ }$根据定理1的(b)部分，我们观察到 $\psi(\cdot)$ $\theta(\cdot)$ β 是满足约束条件 (10) 的 $l_{m,n}$, 1(θ, β),$l_{n,m}$,2(ψ) 的极大化者。因此， $f(\cdot)$ 和 $g(\cdot)$ 的估计量可通过以下方式得到

$$\widehat{f}(x) = \widehat{\theta}(\widehat{\beta}^\tau x)\widehat{\psi}(x)/\lambda, \qquad \widehat{g}(x) = \left\{1 - \widehat{\theta}(\widehat{\beta}^\tau x)\right\} \widehat{\psi}(x)/(1-\lambda).$$

3 计算算法

我们提出一种数值算法，用于计算第2节中所提出的估计量。挑战在于，目标函数 (14) 和 (15) 由 d 维积分构成，直接计算的计算量可能很大。为避免这种多维积分，我们提出一种蒙特卡罗估计程序。

首先，对于数据 $X_1, \ldots, X_n, Y_1, \ldots, Y_m$，我们需要带宽 H_1 和 H_2，以通过以下方式计算$\sim f_n(x)$ 和 $\widetilde{g}_m(x)$

$$\widetilde{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} K_{H_1}(x - X_i), \quad \widetilde{g}_m(x) = \frac{1}{m} \sum_{j=1}^{m} K_{H_2}(x - Y_j).$$

我们采用Wand等人（1994）提出的插件法来获得 H_1 和 H_2，其基于 X_1、.....、 X_n 和

Y_1、……、Y_m，分别。

~接下来，我们从 $f_n(\cdot)$ 中生成 N 个样本 Z_1、……、Z_N，并从 $\widetilde{g}_m(\cdot)$ 中生成 N 个样本 Z_{N+1}、……、Z_{2N}。具体而言，要从 $f_n(\cdot)$ 中生成一个随机变量 Z，可以按以下步骤进行：

- 步骤 1：以相同的概率从 X_1、……、X_n 中抽取一个 X。
- 步骤 2：从 $K(\cdot)$中抽取一个 U。
- 步骤 3：令 $Z = X + H_1 U$。

~我们可以验证，给定 X_1、……、X_n，$Z \sim f_n(x)$。特别地，Z 的条件累积分布函数为

$$\begin{aligned} & P(Z \le x | X_1, \ldots, X_n) = P(X + H_1 \cdot U \le x | X_1, \ldots, X_n) \\ = \; & \frac{1}{n} \sum_{i=1}^{n} P\left(U \le H_1^{-1}(x - X_i) | X_i\right) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{K}(H_1^{-1}(x - X_i)), \end{aligned}$$

~这就是 $f_n(x)$ 对应的条件累积分布函数，其中 $\mathcal{K}(\cdot)$ 表示 $K(\cdot)$ 对应的累积分布函数。类似地，给定 Y_1、……、Y_m，我们可以生成 $Z \sim \widetilde{g}_m(x)$。随后，目标函数 (14) 可估计为

$$\begin{aligned} & \int_{\mathbb{R}^d} \left\{ \frac{\lambda \widetilde{f}_n(x)}{\widehat{\psi}_{n,m}(x)} - \theta(\beta^\tau x) \right\}^2 \widehat{\psi}_{n,m}(x) dx \qquad (16) \\ \approx \; & \frac{1}{N} \lambda \sum_{i=1}^{N} \left\{ \frac{\lambda \widetilde{f}_n(Z_i)}{\widehat{\psi}_{n,m}(Z_i)} - \theta(\beta^\tau Z_i) \right\}^2 + \frac{1}{N}(1-\lambda) \sum_{i=N+1}^{2N} \left\{ \frac{\lambda \widetilde{f}_n(Z_i)}{\widehat{\psi}_{n,m}(Z_i)} - \theta(\beta^\tau Z_i) \right\}^2. \end{aligned}$$

为确保近似精度，我们可以将 N 设为一个较大的值；具体而言，在我们的数值研究中，我们使用了 $N = 10,000$。对于每个 $\beta \in \mathrm{B}$，通过应用标准池相邻违背者算法（PAVA），即可找到使（16）最大化的 $\theta \in \Theta$（Ayer 等人，1955）；具体步骤如下。

• 步骤 1：将 $\{\beta^\tau Z_i\}_{i=1}^{2N}$ 按从小到大的顺序排列，并记为

$$\beta^\tau Z_{(1)} \leq \ldots \leq \beta^\tau Z_{(2N)}.$$

• 步骤 2：因此，目标函数（16）可写为

$$\frac{1}{N}\sum_{i=1}^{2N}\left\{\frac{\lambda\widetilde{f}_n(Z_{(i)})}{\widehat{\psi}_{n,m}(Z_{(i)})} - \theta(\beta^\tau Z_{(i)})\right\}^2 w_{(i)}, \tag{17}$$

~其中，若 $Z_{(i)}$ 来自 $f_n(\cdot)$，则 $w_{(i)} = \lambda$；否则为 $w_{(i)} = 1-\lambda$。

• 步骤 3：可用的池相邻违背者算法（PAVA），例如 `pava()` R 中的函数，可用于数值求解（17）。

最后，（15）中的剖面似然可近似为

$$\frac{1}{N}\lambda\sum_{i=1}^{N}\log\left\{\theta_\beta(\beta^\tau Z_i)\right\} + \frac{1}{N}(1-\lambda)\sum_{i=N+1}^{2N}\log\left\{1-\theta_\beta(\beta^\tau Z_i)\right\}, \tag{18}$$

并可通过 R 中的 optim() 函数轻松求解。

4 渐近性质

本节探讨我们的方法估计量的渐近性质。令 $F_0(\cdot)$、$f_0(\cdot)$、$G_0(\cdot)$、$g_0(\cdot)$、$\theta_0(\cdot)$ 和 β_0 分别表示其对应的量，并令

$$d(\theta_1,\beta_1;\theta_2,\beta_2) = \left[\int\left\{\theta_1(\beta_1^\tau x) - \theta_2(\beta_2^\tau x)\right\}^2 d\Psi_0(x)\right]^{1/2},$$

其中 $\Psi_0(x) = \lambda F_0(x) + (1-\lambda)G_0(x)$。此外，我们记

$$\alpha_n = \max_{H\in\{H_1,H_2\};f\in\{f_0,g_0\}} \int_{\mathbb{R}^d} |K_H * f(x) - f(x)|dx,$$

其中，$K * f(x)$表示函数 K和 f的卷积：$K * f(x) = \int_{\mathbb{R}^d} K(x-y)f(y)dy$。根据核平滑的标准结果，我们有 $\alpha_n \to 0$ 当 $n \to \infty$时成立。特别地，我们有以下引理。

引理1. 设 $H = diag(h_1, \ldots, h_d)$，使得对每个 $i = 1, \ldots, d$，$h_i \to 0$，且 $K(\cdot)$ 满足附录中的条件*1* 。此外，设 $f(x)$ 是定义在 $x \in \mathbb{R}^d$上的概率密度函数，满足

$$\int |D^\alpha f(x)|dx < \infty,$$

其中 $\alpha = (\alpha_1, \ldots, \alpha_d)^\tau$使得 $\alpha_1 + \ldots + \alpha_d = 2$，且 $D^\alpha f = \frac{\partial^{\alpha_1+\cdots+\alpha_d} f}{\partial x_1^{\alpha_1}\cdots\partial x_d^{\alpha_d}}$。那么，

$$\int |K_H * f(x) - f(x)|\, dx = O\left\{\left(\max_{1\le i\le d} h_i\right)^2\right\}.$$

我们有以下定理。

定理2。 假设附录中的条件*1–6* 成立。我们有

(a) $d(\widehat{\theta}, \widehat{\beta}; \theta_0, \beta_0) = O_p(n^{-1/3}) \vee O_p(\sqrt{\alpha_n})$,

(b) $\widehat{\beta} - \beta_0 = O_p(n^{-1/3}) \vee O_p(\sqrt{\alpha_n})$.

$\widehat{}$基于定理2，我们能够建立 f（·）和 $\widehat{g}$（·）的收敛性。

推论1. 假设附录中的条件*1–7* 。

$$\eta_n = \max\{|H_1|,\ |H_2|\},$$

其中 $|H| = h_1 \cdots h_d$ 对于 $H = \text{diag}(h_1, \ldots, h_d)$。假设 $n\eta_n \to \infty$，我们有

$$\begin{aligned} \int \left\{\widehat{f}(x) - f_0(x)\right\}^2 dx &= O_p\left(n^{-2/3}\right) \vee O_p\left(\alpha_n^2\right) \vee O_p\left(1/(n\eta_n)\right) \\ \int \left\{\widehat{g}(x) - g_0(x)\right\}^2 dx &= O_p\left(n^{-2/3}\right) \vee O_p\left(\alpha_n^2\right) \vee O_p\left(1/(n\eta_n)\right). \end{aligned}$$

备注1。考虑以下情形：$H_1 = H_2 = diag(h, \ldots, h)$，并假设引理 *1* 的条件均满足。在此设定下，定理 *2* 中的收敛速率变为

$$O_p\left(n^{-1/3}\right) \vee O_p(\sqrt{\alpha_n}) = O_p\left(n^{-1/3}\right) \vee O_p(h),$$

当我们选择 $O_p\left(n^{-1/3}\right)$ 时，可简化为 $h = O_p\left(n^{-1/3}\right)$。根据定理 *2*，这是我们的方法所能达到的最佳收敛速率。

^类似地，对于 $d \geq 2$，根据推论 l_2，$f(\cdot)$ 和 $\widehat{g}(\cdot)$的收敛速率为

$$\sqrt{O_p\left(n^{-2/3}\right) \vee O_p\left(\alpha_n^2\right) \vee O_p\left(1/(n\eta_n)\right)} = O_p\left(n^{-1/3}\right) \vee O_p(h^2) \vee O_p\left(1/\sqrt{nh^d}\right),$$

最优速率为 $O_p\left\{n^{-2/(4+d)}\right\}$，在 $h = O_p\left\{n^{-1/(4+d)}\right\}$ 时达到 .

5 受试者工作特征和曲线下面积的估计

ROC曲线和曲线下面积已成为许多科学领域广泛使用的统计工具和测量指标，尤其是在医学研究中。在第2节的框架下，组合后的~生物标志物分别为 $U = \beta^\tau X$，$U = \beta^\tau Y$。记 $F_C(u) = P(U \leq u)$和 $G_C(u) =$

$P\left(\widetilde{U} \leq u\right)$。ROC曲线和曲线下面积由Pepe (2003年）给出：

$$\begin{aligned} \mathrm{ROC}_C(s) &= 1 - F_C\left(G_C^{-1}(1-s)\right) \\ \mathrm{AUC}_C &= \int_0^1 \mathrm{ROC}_C(s)ds, \end{aligned}$$

对于 $s \in [0,1]$。此外，当 $\beta = \beta_0$ 时，我们相应地记 F_C,0(u)、G_C,0(u)、ROC$_C$,0(s) 和 AUC$_C$,0；在这种情况下，ROC$_C$,0(s) 是最优ROC曲线（Eguchi and Copas，2002年；McIntosh and Pepe，2002年）。利用我们在第2节中对 β、f(·) 和 g(·) 的估计值，我们可以通过以下方式估计 F_C(·) 和 G_C(·)：

$$\begin{aligned} \widehat{F}_C(u) &= \int I\left(\widehat{\beta}^\tau x \leq u\right)\widehat{f}(x)\,dx, \\ \widehat{G}_C(u) &= \int I\left(\widehat{\beta}^\tau x \leq u\right)\widehat{g}(x)\,dx. \end{aligned}$$

因此 ，相应的ROC曲线估计值和曲线下面积估计值为 y

$$\begin{aligned} \widehat{\mathrm{ROC}}_C(s) &= 1 - \widehat{F}_C\left(\widehat{G}_C^{-1}(1-s)\right) \text{ for each } s \in (0,1), \\ \widehat{\mathrm{AUC}}_C &= \int_0^1 \widehat{\mathrm{ROC}}_C(s)ds. \end{aligned} \tag{19}$$

我们观察到，ROC$_C$(·) 满足以下命题。

命题1。 假设附录中的条件*1* 成立。由 (19) 给出的ROC估计值 $\widehat{\mathrm{ROC}}_C(0) = 0$、$\widehat{\mathrm{ROC}}_C(1) = 1$以及 $\widehat{\mathrm{ROC}}_C(s)$是一个关于 $s \in [0,1]$的非递减凹函数。

以下定理确立了这些估计值的渐近性质。

定理 3。 假设附录中的条件*1*–*6* 。我们有

(a) $\sup_{u\in\mathbb{R}} \left|\widehat{F}_C(u) - F_{C,0}(u)\right| = O_p(n^{-1/3}) \vee O_p(\sqrt{\alpha_n})$,

$$\sup_{u\in\mathbb{R}}\left|\widehat{G}_C(u)-G_{C,0}(u)\right|=O_p(n^{-1/3})\vee O_p(\sqrt{\alpha_n});$$

(b) $\sup_{s\in(0,1)}\left|\widehat{\mathrm{ROC}}_C(s)-\mathrm{ROC}_{C,0}(s)\right|=O_p(n^{-1/3})\vee O_p(\sqrt{\alpha_n});$

(c) $\widehat{\mathrm{AUC}}_C=\mathrm{AUC}_{C,0}+O_p\left(n^{-1/3}\right)\vee O_p(\sqrt{\alpha_n}).$

根据定理 3 以及备注 1中的讨论，所有这些估计量的最佳收敛速率为 $O_p\left(n^{-1/3}\right)$，当 $h=O_p\left(n^{-1/3}\right)$ 时达到。

6 模拟研究

6.1 设置

在我们的数值研究中，我们将我们的方法（称为“Our”）与三种具有竞争力的方法进行比较：(1) Chen等人（2016年）提出的经验似然方法（称为“Empirical”）；(2) Qin and Zhang （2010）提出的指数倾斜方法（称为“Exp Tilting”）；(3) Ma and Huang （2007）提出的基于最大化曲线下面积建立的方法（称为“MH”）。

我们考虑两个模拟示例：例1和例2。对于例1，我们设定 $d=2$。令 $f_1(\cdot)$ $f_2(\cdot)$ 和 $f_3(\cdot)$ 分别为对数正态分布 lognormal(0,1)、lognormal(1,1) 和 lognormal(4,1) 的概率密度函数。我们模拟 X_1，...，X_n 为独立同分布，其满足

$$X_i\sim f(x_1,x_2)=\rho f_1(x_1)f_1(x_2)+(1-\rho)f_2(x_1)f_3(x_2);$$

并模拟 $Y_1,\ldots,Y_m$，使其独立同分布，且

$$Y_i\sim g(x_1,x_2)=(1-\rho)f_1(x_1)f_1(x_2)+\rho f_2(x_1)f_3(x_2),$$

其中，我们的方法在数值研究中考察了 ρ 的不同取值。请注意，对数正态分布(μ, σ)的概率密度函数为

$$\frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\log x - \mu)^2}{2\sigma^2}\right\}.$$

因此， $U(x_1, x_2) \equiv \log\left\{\frac{f_1}{f_2}(x_1)f_1(x_2)\,(x_1)f_3(x_2)\right\}$具有如下结构:

$$U(x_1, x_2) = \log\left\{\frac{f_1(x_1)f_1(x_2)}{f_2(x_1)f_3(x_2)}\right\} = \alpha + \beta_1 \log x_1 + \beta_2 \log x_2. \tag{20}$$

此外,

$$\frac{f(x_1, x_2)}{g(x_1, x_2)} = \frac{\rho \exp(U) + (1-\rho)}{(1-\rho)\exp(U) + \rho}. \tag{21}$$

在实际应用中， $f_1(x_1)f_1(x_2)$ 和 $f_2(x_1)f_3(x_2)$ 可能分别表示基于真实值的患病者和健康受试者的生物标志物联合概率密度函数。 $1-\rho$表示识别出的患病组中实际为健康者的比例；同样，也表示识别出的健康组中实际为患病者的比例。因此， $\rho = 1$对应于所有患病者和健康受试者均被正确识别的理想情形。在这种情况下，$\log\left(f(x_1, x_2)/g(x_1, x_2)\right) = U(x_1, x_2)$，因此我们的方法和所有竞争方法的模型假设均完全满足。类似地， $\rho \in (0.5, 1)$ 对应于患病组和健康组中均有比例为 $1-\rho$ 的受试者被误分类的情形。根据（20）和（21），“ 我们的方法 ”、“ 经验 ” 和 “Metropolis–Hastings 算法 ” 方法的模型假设均满足，而“指数倾斜”方法的模型假设则不满足。

对于示例2，我们考虑 $d = 3$。令 $g_1(\cdot) \sim N(0, 1)$， $g_2(\cdot) \sim$伽马分布$(2.5, 4)$， $g_3(\cdot) \sim$

$N(1, 1)$， $g_4(\cdot) \sim N(4.5, 1)$， $g_5(\cdot)$ ～伽马分布$(2, 4)$。我们通过以下方式模拟 X_1，……， X_n，以及 Y_1，……， Y_m

$$\begin{aligned} X_i &\sim f(x_1, x_2, x_3) = \rho g_1(x_1)g_1(x_2)g_2(x_3) + (1-\rho)g_3(x_1)g_4(x_2)g_5(x_3) \\ Y_i &\sim g(x_1, x_2, x_3) = (1-\rho)g_1(x_1)g_1(x_2)g_2(x_3) + \rho g_3(x_1)g_4(x_2)g_5(x_3). \end{aligned}$$

与例1类似， $g_1(x_1)g_1(x_2)g_2(x_3)$ 和 $g_3(x_1)g_4(x_2)g_5(x_3)$ 可分别表示患病者和健康受试者生物标志物的联合概率密度函数，并满足

$$U(x_1, x_2, x_3) = \log\left\{\frac{g_1(x_1)g_1(x_2)g_2(x_3)}{g_3(x_1)g_4(x_2)g_5(x_3)}\right\} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \log x_3.$$

数量 $1-\rho$ 表示识别出的患病组中实际健康者所占的比例，同样也表示识别出的健康组中实际患病者所占的比例。

对于这两个例子，我们考虑以下设定组合： $\rho = 1$、0.9、0.8； $n = m = 300$ 和 $n = m = 600$。

6.2 β 估计值的结果

在本节中，我们比较不同方法得出的 β 估计值。为便于比较，我们对 $\|\beta\|_2 = 1$进行正则化，其中 $\|\cdot\|_2$ 表示欧几里得空间中的 l_2 范数。表1 和2 汇总了例1和例2中不同方法在1000次重复试验下的结果；所有报告值均乘以1000，以便更清晰地呈现。从两张表中可以看出，对于这两个例子，我们的方法在不同样本量和模型设定错误程度下，始终取得最低或接近最低的均方误差（MSE）和标准差（SD）。即使 ρ 较小，其表现仍然稳定，表明对模型错设具有很强的稳健性

。经验方法和指数倾斜方法的准确度中等，但在模型错设程度较强时，表现会下降。MH方法的表现通常最差，在模型错设程度较强且样本量较小时，其变异性较大，均方误差也较高。这表明，在此类情形下，这些方法的稳健性和效率有限。

表1：例1的 β 估计值比较。所有报告值均已乘以1000

.

			Our			经验			指数倾斜			MH		
$n = m$	ρ	β	Bias	SD	MSE	Bias	SD	MSE	Bias	SD	MSE	Bias	SD	MSE
	1.0	β_1	-2.87	60.97	3.73	6.70	99.24	9.90	7.81	82.64	6.89	6.40	93.49	8.78
	1.0	β_2	2.77	15.67	0.25	3.75	25.11	0.64	1.82	20.84	0.44	3.23	24.11	0.60
300	0.9	β_1	-13.57	85.40	7.48	2.39	151.81	23.05	0.48	116.93	13.67	16.45	302.67	91.88
	0.9	β_2	7.56	23.63	0.62	12.24	41.57	1.88	7.43	30.45	0.98	48.24	86.44	9.80
	0.8	β_1	-20.65	120.48	14.94	4.43	221.76	49.20	3.54	162.80	26.52	8.62	371.52	138.10
	0.8	β_2	13.58	34.81	1.40	26.93	67.40	5.27	13.85	43.68	2.10	78.52	111.42	18.58
	1.0	β_1	-0.05	47.03	2.21	5.85	73.38	5.42	1.14	55.87	3.12	-0.66	63.10	3.98
	1.0	β_2	1.22	11.74	0.14	1.50	18.08	0.33	1.43	14.13	0.20	2.36	16.39	0.27
600	0.9	β_1	-7.77	64.70	4.25	-0.81	112.50	12.66	-1.21	80.78	6.53	9.77	249.32	62.25
	0.9	β_2	4.29	17.30	0.32	7.22	30.24	0.97	3.90	21.22	0.47	32.59	68.70	5.78
	0.8	β_1	-15.07	90.58	8.43	4.37	156.60	24.54	1.71	112.18	12.59	14.58	317.64	101.11
	0.8	β_2	8.48	25.43	0.72	12.55	42.40	2.00	6.55	30.43	0.97	54.23	91.37	11.29

表 2：示例2的 β 估计值比较。所有报告的数值均已乘以 1000。

			Our			经验			指数倾斜			MH		
$n=m$	ρ	β	Bias	SD	MSE	Bias	SD	MSE	Bias	SD	MSE	Bias	SD	MSE
	1.0	β_1	3.56	78.60	6.19	8.71	119.61	14.38	0.81	102.21	10.45	10.72	112.00	12.66
	1.0	β_2	7.84	22.33	0.56	20.58	37.46	1.83	15.24	31.56	1.23	17.95	34.75	1.53
	1.0	β_3	-8.74	108.79	11.91	6.77	161.35	26.08	-6.66	140.30	19.73	4.88	155.68	24.26
	0.9	β_1	-22.28	107.80	12.12	5.86	181.67	33.04	2.74	124.59	15.53	45.81	324.61	107.47
300	0.9	β_2	24.89	37.56	2.03	48.89	63.44	6.42	22.13	38.03	1.94	145.94	109.66	33.32
	0.9	β_3	12.49	147.44	21.89	7.36	237.52	56.47	-0.84	163.97	26.89	-22.34	408.47	167.35
	0.8	β_1	-32.01	143.08	21.50	21.06	242.85	59.42	6.79	165.61	27.47	36.85	371.64	139.47
	0.8	β_2	44.25	56.28	5.13	83.17	88.32	14.72	39.50	56.59	4.76	190.87	125.83	52.26
	0.8	β_3	15.86	204.59	42.11	34.12	296.67	89.17	0.46	217.60	47.35	-22.83	446.53	199.91
	1.0	β_1	3.51	60.96	3.73	5.71	88.66	7.89	3.25	70.81	5.03	4.16	80.65	6.52
	1.0	β_2	4.13	16.12	0.28	10.11	25.40	0.75	6.11	19.33	0.41	8.78	23.52	0.63
	1.0	β_3	-5.42	81.67	6.70	2.99	113.18	12.82	-6.49	95.87	9.23	-3.52	111.59	12.46
	0.9	β_1	-19.27	79.12	6.63	13.88	134.12	18.18	6.90	90.76	8.28	29.96	297.77	89.56
600	0.9	β_2	16.49	25.40	0.92	25.83	43.21	2.53	10.67	25.37	0.76	119.55	97.42	23.78
	0.9	β_3	11.01	116.88	13.78	15.26	178.70	32.17	-3.03	123.76	15.33	-18.26	368.05	135.80
	0.8	β_1	-32.89	108.13	12.77	32.40	184.40	35.05	10.30	122.86	15.20	37.55	339.43	116.62
	0.8	β_2	29.58	39.22	2.41	44.62	63.19	5.98	19.38	35.22	1.62	161.16	110.67	38.22
	0.8	β_3	15.34	156.65	24.77	8.57	240.24	57.79	-8.25	164.28	27.06	-16.49	421.61	178.03

6.3 受试者工作特征和曲线下面积估计值的结果

在本节中，我们将我们的方法所得的受试者工作特征和曲线下面积估计值与其他方法的结果进行比较。对于受试者工作特征估计，我们以 1000 次重复试验中受试者工作特征 $\widehat{C}(\cdot)$ 与受试者工作特征 $C_0(\cdot)$ 之间的平均 L_2 距离作为比较标准；具体而言，该 L_2 距离定义为：

定义如下：

$$L_2\left(\widehat{\mathrm{ROC}}_C, \mathrm{ROC}_{\mathrm{C},0}\right) = \left[\int_0^1 \left\{\widehat{\mathrm{ROC}}_C(s) - \mathrm{ROC}_{\mathrm{C},0}(\mathrm{s})\right\}^2 ds\right]^{1/2}.$$

对于AUC估计值，我们采用1000次重复试验的相对偏差（RB）和均方误差（MSE）作为比较标准。百分比表示的RB定义如下：

$$RB(\%) = \frac{\widehat{\mathrm{AUC}}_C - \mathrm{AUC}_{C,0}}{\mathrm{AUC}_{C,0}} \times 100.$$

表3 展示了 L_2 例1中不同样本量和误分类率下ROC估计的距离，以及AUC估计的RB（%）和MSE。类似地，表4展示了例2的相应结果。根据这些表格，我们有以下观察。

我们的方法在所有情形下始终取得最低或接近最低的 L_2 距离，表明ROC曲线估计更准确。经验方法总体表现第二好，但当 ρ 降低时，其 L_2 误差比我们的方法更大。指数倾斜法的 L_2 距离往往较大，尤其是在 ρ 值较低时；这符合预期，因为较低的 ρ 表明该方法的模型错设程度更高。MH方法的表现与经验方法相当。

在AUC估计方面，我们的方法在大多数情况下均取得最小的均方误差和绝对相对偏差，表明其在所有方法中偏差最小、精确度最高。MH方法的偏差也较小，与我们的方法相当，但其均方误差往往略大，说明AUC估计值的精确度较低。相比之下，经验法和指数倾斜法的相对偏差较大，且均方误差通常高于或与MH方法相当。

总之，无论在ROC曲线还是AUC估计准确度方面，我们的方法都优于其他方法，即使在存在中等程度错误分类且样本量较小时，也能保持较低的偏差和误差。经验法和MH方法的表现尚可，但对错误分类误差的稳健性较差。指数倾斜法的误差和偏差始终最大，表明在存在错误分类组的设置中，该方法可能不太适用。

表3：例1：ROC估计的 L_2 距离（$\times 1000$），以及AUC估计。

		Our			经验			指数倾斜			MH		
$n=m$	ρ	L_2	RB	MSE	L_2	RB	MSE	L_2	RB	MSE	L_2	RB	MSE
300	1.0	6.97	-0.17	<0.01	4.86	0.03	<0.01	3.32	-0.02	<0.01	4.91	0.00	<0.01
	0.9	27.53	-0.09	0.17	33.50	1.20	0.27	49.73	0.90	0.24	32.59	0.11	0.19
	0.8	26.77	0.22	0.31	34.09	2.21	0.61	44.32	1.14	0.42	32.33	0.41	0.35
600	1.0	5.68	-0.14	<0.01	3.65	0.01	<0.01	2.31	-0.02	<0.01	3.75	-0.01	<0.01
	0.9	20.24	-0.16	0.09	23.62	0.73	0.13	47.02	0.85	0.15	23.31	-0.09	0.10
	0.8	19.50	0.23	0.16	24.21	1.56	0.31	41.00	1.19	0.26	23.08	0.27	0.18

表4：示例2：ROC估计的 L_2 距离（$\times 1000$），以及AUC估计的相对偏差（%）和均方误差（$\times 1000$）。AUC估计。

		Our			经验			指数倾斜			MH		
$n=m$	ρ	L_2	RB	MSE	L_2	RB	MSE	L_2	RB	MSE	L_2	RB	MSE
300	1.0	2.52	-0.06	<0.01	2.44	-0.02	<0.01	1.87	-0.03	<0.01	2.37	-0.02	<0.01
	0.9	28.54	0.47	0.20	35.73	1.53	0.35	48.13	1.42	0.34	33.82	0.52	0.22
	0.8	28.49	0.80	0.39	36.79	2.57	0.75	43.40	1.69	0.55	33.79	0.89	0.42
600	1.0	2.13	-0.06	<0.01	1.92	-0.02	<0.01	1.31	-0.04	<0.01	1.84	-0.03	<0.01
	0.9	20.94	0.26	0.10	24.57	0.96	0.16	44.99	1.32	0.23	24.12	0.19	0.10
	0.8	20.03	0.48	0.18	25.03	1.62	0.32	38.94	1.49	0.31	23.60	0.39	0.18

7 应用

胰腺癌是最致命的癌症之一，确诊后的五年生存率不足10%。然而，早期发现能显著提高生存机会。不幸的是，胰腺癌往往在发展到晚期并发生转移之前都没有症状。因此，能够在早期识别胰腺癌的可靠诊断检测具有重要的临床价值。

在Debernardi 等人（2020年）中，作者从三组患者中收集了一系列尿液生物标志物：健康对照组；患有慢性胰腺炎等非癌性胰腺疾病的患者；以及胰腺导管腺癌患者。这些组的样本量分别为183、143和83。检测的尿液蛋白质包括 LYVE1、REG1B 和 TFF1。我们将 LYVE1、REG1B 和 TFF1 的检测结果结合起来。为此，我们采用以下半参数单调密度比模型：

$$\frac{f(x)}{g(x)} = \phi\left(\beta_1(100 \times \text{LYVE1}) + \beta_2(\text{REG1B}) + \beta_3(\text{TFF1})\right),$$

其中 multip　将 LYVE1 乘以 100，使其量级与其他生物标志物保持一致　　.

我们使用同一机构——英国伦敦的 Barts 胰腺组织库（BRTB）收集的数据。我们将患有非癌性胰腺疾病的患者视为健康人，因此非患病组样本包括健康对照组和非癌性胰腺疾病组，即 $n = 83$， $m = 326$。我们报告基于 300 个自助法估计值计算得出的系数估计值及其对应的标准误（SE）；为便于比较，我们对 $\|\beta\|_2 = 1$ 进行了正则化。结果汇总于表5。

我们观察到，四种方法得到的估计值略有不同；我们的方法与“经验”方法之间的差异最小。与其他方法相比，我们的方法所得的标准误更小。

与模拟示例不同，我们无法获得这一真实数据实例中 β 估计值的偏差和均方误差。然而，根据我们在模拟研究中的观察，我们推测，非癌性胰腺疾病患者的数据会导致指数倾斜模型发生模型错设。

表5：系数的估计值及其标准误 SE($\times 10$)

	Our		经验		指数倾斜		MH	
参数	估计值	SE	估计值	SE	估计值	SE	估计值	SE
β_1	0.90	1.01	0.90	1.30	0.74	1.50	0.84	2.04
β_2	0.43	1.26	0.44	1.82	0.67	1.60	0.54	2.57
β_3	0.02	0.52	0.02	0.54	0.02	0.63	0.05	0.73

基于 β 估计值，我们使用四种方法构建ROC曲线，如图1所示，并在表6中汇总相应的AUC估计值。我们的方法、经验方法和指数倾斜方法得到的ROC曲线和AUC值相近，表明总体上具有一致性。MH方法在ROC曲线和AUC估计值方面均表现出略大的偏差。然而，与模拟研究不同，由于真实数据中缺乏真实值，我们无法根据真实值评估这些估计值的准确性。

表6：四种曲线下面积估计方法的估计值和标准误（$\times 1000$）。

Our		经验		指数倾斜		MH	
估计值	SE	估计值	SE	估计值	SE	估计值	SE
0.84	20.30	0.86	18.07	0.81	23.53	0.85	21.37

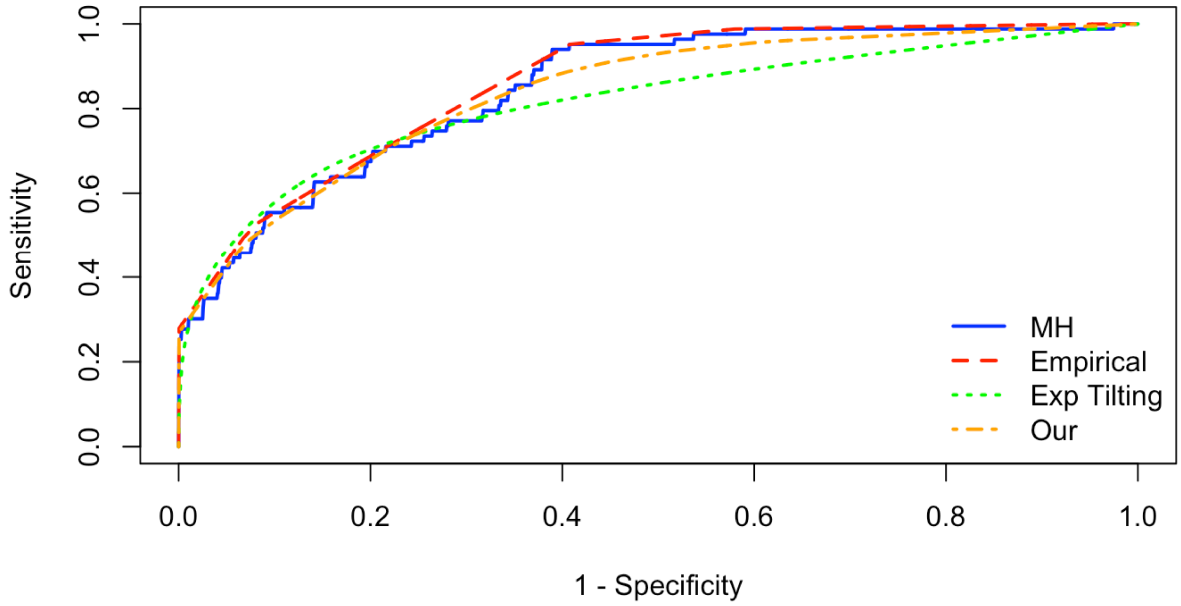

图1：ROC估计

8 讨论

本文介绍了一种在（2）所定义的半参数密度比模型下，用于结合多项诊断测试的最大平滑似然方法。所提出的方法兼顾了非参数成分的灵活性和平滑性，从而提高了参数估计的准确性，并增强了对模型设定错误的稳健性。

我们开发了一种高效算法，确立了参数估计量的渐近性质，并推导出ROC曲线和曲线下面积的估计量及其理论保证。通过数值研究，我们证明，所提出的方法能够持续为组合系数和ROC/AUC指标提供更准确、更稳定的估计，尤其是在模型设定错误的情况下。

尽管所提出的方法具有诸多优势，其性能仍可通过改进

核平滑的带宽选择来提升。此外，当生物标志物数量 d 较大时，该方法可能面临计算方面的挑战，而且目前的收敛速率可能并非最优。未来的研究可着眼于实现带宽选择自动化、提升方法在高维设置下的可扩展性，并建立更强的理论保证。

附录：正则性条件

为建立渐近结果，我们施加以下正则性条件。这些条件未必是可能的最弱条件。

条件1。 $K(\cdot)$ 是具有紧支撑的对称概率密度函数，支撑位于 $\mathbb{R}^d$，且 $\sup_{t\in\mathbb{R}^d} K(t) \cdot 1$，其中“.”表示“至多相差一个普适常数”。

条件2。 $\beta \in \{1\} \times \mathrm{B}_{-1}$ 和 B_{-1} 是 R^{d-1} 的紧子空间

Co条件3。 $H_1 = \mathrm{diag}(h_{1,1}, \ldots, h_{1,d})$， $H_2 = \mathrm{diag}(h_{2,1}, \ldots, h_{2,d})$。它们满足：

- 对每个 $s = 1$， 2； $t = 1, \ldots, d$, $h_{s,t} \to 0$ 当 $n \to \infty$ 时。
- 对于任意 s_1、 $s_2 \in \{1, 2\}$、 t_1、 $t_2 \in \{1, \ldots, d\}$,

$$0 < \liminf_{n\to\infty} h_{s_1,t_1}/h_{s_2,t_2} \le \limsup_{n\to\infty} h_{s_1,t_1}/h_{s_2,t_2} < C$$

其中 C 为普适常数。

条件4。对于 β_1、 $\beta_2 \in \mathrm{B}$ 和 $u \in \mathbb{R}^d$，存在一个普适常数 C，不依赖于 u，使得

$$\int \left| I\left(\beta_1^\tau x \le u\right) - I\left(\beta_2^\tau x \le u\right) \right| dP(x) \le C \left\| \beta_1 - \beta_2 \right\|_1,$$

对于 $P = F_0$、 G_0 和 $\mathcal{K}$。

条件5。存在一个 $\eta_0 > 0$，使得

$$\inf_{\beta:\|\beta-\beta_0\|_2\leq\eta_0} \lambda_2 \left[var\left\{E\left(X|\beta^\tau X\right) - X\right\}\right] > 0,$$

其中， $\lambda_2(A)$ 表示矩阵 A的第二小特征值。

条件6. $\theta_0(\cdot)$ 二阶连续可微，且二阶导数有界；并且存在普适常数 $0 < c_1 < c_2 < 1$，使得对于任意 $x \in \mathbb{R}^d$, $c_1 < \theta_0(\beta_0^\tau x) < c_2$。此外，对于任意 $\beta \in \mathrm{B}$，以及任意满足 $\theta(\beta^\tau x) = \theta_0(\beta_0^\tau x)$ 对所有 x 均成立的单调递增函数 $\theta(\cdot)$，则 $\beta = \beta_0$ and $\theta(\cdot) = \theta_0(\cdot)$。

条件7. $f_0(x)$ 和 $g_0(x)$ 均有界，且对于$x \in \mathbb{R}^d$，二阶导数有界。

注2. 在条件 *4*下，我们有

$$\int \left| I\left(\widetilde{\beta}_1^\tau z \leq u\right) - I\left(\widetilde{\beta}_2^\tau z \leq u\right)\right| dP(z) \leq C \left\|\widetilde{\beta}_1 - \widetilde{\beta}_2\right\|_1,$$

$\widetilde{\ }$ $\widetilde{\ }$其中 $\beta_1 = (\beta_1^T, \beta_{1,\ 1}^T)^T$, $\beta_2 = (\beta_2^T, \beta_{2,\ 1}^T)^T$，其中 $\beta_1, \beta_{1,\ 1}, \beta_2, \beta_{2,\ 1} \in \mathrm{B}$，$z = (x^T, t^T)^T \in \mathbb{R}^{2d}$，以及$P(z) = F_0(x)\mathcal{K}(t)$ 或 $P(z) = G_0(x)\mathcal{K}(t)$。

证明。该备注由以下内容得出：

$$
\begin{aligned}
& \int \left| I\left(\widetilde{\beta}_1^\tau z \le u\right) - I\left(\widetilde{\beta}_2^\tau z \le u\right) \right| \, dP(z) \\
= & \int\int \left| I\left(\beta_1^\tau x \le u - \beta_{1,1}^\tau t\right) - I\left(\beta_2^\tau x \le u - \beta_{2,1}^\tau t\right) \right| \, dF_0(x) d\mathcal{K}(t) \\
\le & \int\int \left| I\left(\beta_1^\tau x \le u - \beta_{1,1}^\tau t\right) - I\left(\beta_2^\tau x \le u - \beta_{1,1}^\tau t\right) \right| \, dF_0(x) d\mathcal{K}(t) \\
& + \int\int \left| I\left(\beta_2^\tau x \le u - \beta_{1,1}^\tau t\right) - I\left(\beta_2^\tau x \le u - \beta_{2,1}^\tau t\right) \right| \, dF_0(x) d\mathcal{K}(t) \\
\le & C\|\beta_1 - \beta_2\|_1 + \int\int \left| I\left(\beta_{1,1}^\tau t \le u - \beta_2^\tau x\right) - I\left(\beta_{2,1}^\tau t \le u - \beta_2^\tau x\right) \right| \, d\mathcal{K}(t) dF_0(x) \\
\le & C\|\beta_1 - \beta_2\|_1 + C\|\beta_{1,1} - \beta_{2,1}\|_1 = C\|\widetilde{\beta}_1 - \widetilde{\beta}_2\|_1.
\end{aligned}
$$

补充材料

补充材料包含第2、4和5节理论结果的技术细节。

数据可用性

真实的胰腺　　癌症数据在第7 节中进行了分析，可从以下网址下载：

https://www.kaggle.com/datasets/johnjdavisiv/urinary-biomarkers-for-pancreatic-cancer

参考文献

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., 和 Silverman, E. (1955)，“不完整信息抽样的经验分布函数”，《数理统计年刊》，26，641–647。

Barreno, M., Cardenas, A. A., 和 Tygar, J. D. (2008年), “分类器组合的最优ROC曲线”, 载于《神经信息处理系统进展》（NIPS），第20卷，第57–64页。

Chen, B., Li, P., Qin, J., 等（2016年），《使用单调密度比模型寻找多项诊断测试的渐近最优组合》，Journal of the American Statistical Association，111，861–874。

Copas, J. 和 Corbett, P.（2002年），《逻辑回归中受试者工作特征曲线的高估》，Biometrika，89，315–331。

Debernardi, S., O’Brien, H., Algahmdi, A. S., Malats, N., Stewart, G. D., Pljeˇ sa-Ercegovac, M., Costello, E., Greenhalf, W., Saad, A., Roberts, R., 等（2020年），《尿液生物标志物组合与PancRISK评分相结合以实现胰腺癌早期检测：一项病例对照研究》，PLoS Medicine，17，e1003489。

Eggermont, P. P. B. 和 LaRiccia, V. N.（1995a），《最大平滑似然密度估计》，《非参数统计》，4，211–222。

—（1995b），《逆问题的最大平滑似然密度估计》，《统计学年鉴》，23，199–220。

Eggermont, P. P. B. 和 LaRiccia, V. N.（2000），《光滑单调及单峰密度的最大似然估计》，《统计学年鉴》，28，922–947。

Eggermont, P. P. B. 和 LaRiccia, V. N.（2001年），《最大惩罚似然估计》，纽约：Springer。

Eguchi, S. 和 Copas, J.（2002年），“一类逻辑型判别函数”，Biometrika，89，1–22。

Ghosal, S.（2025年），“方法学假设和协变量对ROC分析中临界值估计的影响”，《生物计量学杂志》，67，e70053。

Groeneboom, P. (2014),“区间删失模型的最大平滑似然估计量”，《统计学年鉴》,42, 2092–2137。

Groeneboom, P. 和 Jongbloed, G. (2010),“广义连续保序回归”，Statis- tics & ProbabilityLetters, 80, 248–253。

Groeneboom, P., Jongbloed, G. 和 Witte, B. I. (2010),“当前状态模型中的最大平滑似然估计和平滑最大似然估计”，《统计学年鉴》, 38, 352–387。

Huang, Y. 和 Sanda, M. G. (2022),“用于约束分类的线性生物标志物组合”，《统计学年鉴》, 50, 2793–2815。

Kim, Y., Yeo, I., Huh, I., Kim, J., Han, D., Jang, J. 和 Kim, Y. (2021),“用于胰腺癌诊断的蛋白质多标志物组的开发与多重验证”，《临床癌症研究》, 27, 2236–2245，列出了全部七位作者；前三位（Kim, Y.; Yeo, I.; Huh, I.）贡献相同。

Kim, Y.-S., Jang, M.-K., Park, C.-Y., Song, H.-J. 和 Kim, J.-D.（2013年），“利用逻辑回归探索卵巢癌早期筛查的多种生物标志物组合”，《国际生物科学与生物技术杂志》，5，67–76。

Kodosaki, E., Watkins, W. J., Loveless, S., Kreft, K. L., Richards, A., Anderson, V., Hurler, L., Robertson, N. P., Zelek, W. M., Tallantyre, E. C. 等（2024年），《组合蛋白质生物标志物预测多发性硬化的诊断和结局》，《神经炎症杂志》，21，52。

Levine, M., Hunter, D. R., 和 Chauveau, D.（2011），《多元混合模型的最大平滑似然》，Biometrika，98，403–416。

Liu, D. 和 Zhou, X.-H.（2013），《利用协变量调整进行生物标志物组合的ROC分析》，《学术放射学》，20，874–882。

Ma, S. 和 Huang, J.（2007），《使用ROC组合多个标志物进行分类》，《生物计量学》，63，751–757。

McIntosh, M. 和 Pepe, M.（2002），《组合多项筛查试验：风险评分的最优性》，《生物计量学》，58，657–664。

Pepe, M. S.（2003），《用于分类和预测的医学检验统计评估》，纽约：牛津大学出版社。

Qin, J. 和 Zhang, B.（2010），《用于筛查的多项诊断测试的最佳组合》，《医学统计学》，29，2905–2919。

Su, J. 和 Liu, J.（1993），《多个诊断标志物的线性组合》，《美国统计学会杂志》，88，1350–1355。

Wand, M. P., Jones, M. C., 等（1994），《多变量插件带宽选择》，Computa- tional Statistics，9，97–116。

Yu, T., Li, P. 和 Qin, J.（2017），《具有似然比排序的双样本问题中的密度估计》，Biometrika，104，141–152。

—（2019），《已知混合比例的混合模型中的最大平滑似然分量密度估计》，《电子统计学期刊》，13，4035–4078。