arXiv:2601.14631v1 [stat.ML] 21 Jan 2026

随机缺失概念下结合间隔置信度与 Aranda–Ordaz 函数的半监督混合模型

1st Jinyang Liao新南威尔士大学
数学与统计学院澳大利亚悉尼
ji
nyang.liao@student.unsw.edu.au

2nd ZiyangLyu新南威尔士大学数学与统计学院澳大利亚悉尼
ziyang.lyu@unsw.edu.au

摘要——**本文提出了一个针对随机缺失（MAR）机制下高斯混合建模的半监督学习框架。该方法通过将缺失概率建模为分类不确定性的函数，显式地参数化缺失标签机制。为量化分类不确定性，我们引入间隔置信度，并结合 Aranda–Ordaz（AO）链接函数，以灵活刻画不确定性与缺失概率之间的非对称关系。基于这一建模方式，我们提出了一种高效的期望条件最大化（ECM）算法，联合估计高斯混合模型（GMM）和缺失机制中的所有参数，随后利用拟合得到的混合模型所导出的贝叶斯分类器填补缺失标签。该方法有效减轻了忽略缺失机制所导致的偏差，同时增强了半监督学习的稳健性。所得的不确定性感知框架在缺失标签比例较高的现实随机缺失情景中，能够实现可靠的分类性能。**

索引词—**随机缺失机制、AO 链接函数、间隔置信度、GMM、期望条件最大化算法**

I. 引言

现代机器学习应用经常需要处理部分标注的数据集，其中只有少部分数据具有标签，而大多数数据仍未标注。根据 [1]–[4], , 当为所有数据标注的成本高昂或难以实现时，通常会出现这种情况数据点。为了建立一个严谨的框架来推理此类不完整数据， [5]提出了广泛使用的缺失机制分类：完全随机缺失（MCAR）、随机缺失（MAR）和非随机缺失（MNAR）。随后， [6], [7]阐明了 Rubin 分类法，强调 MCAR 与 MAR 在概念和推断上的区别，并指出一个关键难题：实践中的标签通常并非完全随机缺失。在实践中，可在 MAR 框架下借助 [8]–[10],刻画缺失情况；该方法允许缺失依赖于分类不确定性等可观测特征，但不依赖于不可观测标签。这一特性影响深远，因为 [11]–[13] 的证据表明，忽视这种结构会导致参数估计产生偏差，并削弱判别和校准方面的分类性能。因此，在半监督情境下，明确建模标签缺失机制对于准确估计和稳健的预测性能至关重要。在 MAR假设下， [14], [15] 将缺失对可观测特征的依赖纳入模型。本文以 [16]的有限混合建模框架为基础，并遵循 [14], [17]以及 [11], 对半监督学习的处理，在双分量高斯混合建模框架下研究 MAR

并在有限混合模型设定中提出一种新的 MAR 设定。

在对随机缺失机制进行建模时，我们引入间隔置信度这一新概念，用以量化每个观测值的分类不确定性 [18]–[20]。与以往依赖香农熵的研究相比， [14], [17],间隔置信度在保留强判别能力的同时，提供了效率显著更高的近似度量。我们进一步证明，在大多数实际情形下，香农熵可由间隔置信度的平方很好地近似，从而大幅节省计算成本，且模型性能几乎不会损失。此外，以往的方法通常采用 logit 链接函数对随机缺失机制进行建模。然而，在非高斯混合等复杂情形下，logit 链接函数的鲁棒性有限，因为它强制要求响应在决策边界两侧呈对称分布，因此难以灵活刻画高度偏斜或集中于不确定区域的缺失机制。相比之下，Aranda–Ordaz 链接函数通过引入形状参数 λ对 logit 函数进行了推广，从而放宽了逻辑函数 [21],[22] 固有的对称性。通过调整 λ，该链接函数允许缺失概率以非对称方式依赖于基于间隔的不确定性 [22]–[24]。这一选择显著提升了所提出的 MAR 感知混合模型的稳定性和适应性。

在实验部分，我们通过受控模拟和真实数据实验评估所提出的框架。在模拟部分，我们采用基于间隔的随机缺失机制移除标签，从而能够将所提出的 ECM–AO 与忽略缺失机制的逻辑回归基线进行比较。为评估该方法在超越理想化设定时的稳健性，我们还考察了非高斯混合模型和 MAGIC Gamma Telescope 数据集。这些实验评估了 ECM–AO 在分布设定错误和现实标签稀疏性下的表现。

II. 方法论

A. 高斯混合模型

设 $\boldsymbol{Y}$为一个来自 K 分量高斯混合模型的 d 维随机向量。每个分量 k ($k = 1, \ldots\ldots, K$) 都对应一个先验概率 π_k，其中 $\pi_k \geq 0$ 且 $\sum_{k=1}^{K} \pi_k = 1$。在被分配到分量 k 的条件下， $\boldsymbol{Y}$ 的分布为多元高斯分布，其密度为 $\mathcal{N}(\boldsymbol{y};\ \boldsymbol{\mu}_k,\ \boldsymbol{\Sigma}_k)$，其中 $\boldsymbol{\mu}_k \in \mathbb{R}^d$ 表示第 k 个分量的均值向量， $\boldsymbol{\Sigma}_k \in \mathbb{R}^{d\times d}$ 表示其协方差矩阵。因此， $\boldsymbol{Y}$ 的边际密度为

$$f(\boldsymbol{y}_j;\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{y}_j;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k),$$

完整参数向量为

$$\boldsymbol{\theta} = \left\{\pi_1,\ldots,\pi_{K-1},\ \boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K,\ \boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_K\right\}.$$

对于观测 $\boldsymbol{y}_j$，贝叶斯规则将其分配给类别 k，如果 $\pi_k \mathcal{N}\left(\boldsymbol{y}_j \mid \boldsymbol{\mu}_k\ \boldsymbol{\Sigma}_k\right)/\sum_{j=1}^{K} \pi_j \mathcal{N}\left(\boldsymbol{y}_j \mid \boldsymbol{\mu}_j\ \boldsymbol{\Sigma}_j\right)$ $\boldsymbol{y}_j$ $=$ $\arg\max$，其中 $\boldsymbol{y}_j$ ，表示给定 $\boldsymbol{Y} = \boldsymbol{y}_j$ 时观测 $\boldsymbol{y}_j$ 属于类别 k 的后验概率。

考虑训练集 $\{\boldsymbol{y}_j, m_j, z_j\}_{j=1}^{n}$，其中 $\boldsymbol{y}_j \in \mathbb{R}^d$ 表示特征向量， $z_j \in \{1,\ldots,K\}$ 是类别标签， m_j 是缺失指示变量：当观测到 z_j 时取值为 $m_j = 0$ ，否则取值为 $m_j = 1$ 。

为便于构造似然，我们定义指示变量 z_{ij} ，使得当观测 $\boldsymbol{y}_j$ 属于分量 i 时取值为 $z_{ij} = 1$ ，否则取值为 $z_{ij} = 0$，其中 $i = 1, \ldots\ldots, K$，且 $j = 1, \ldots\ldots, n$。遵循 [14],的框架，观测数据对数似然可分解为有标签样本和无标签样本的贡献，表示如下：

$$\log L_C(\boldsymbol{\theta}) = \sum_{j=1}^{n}(1-m_j)\sum_{i=1}^{K} z_{ij}\ \log\left\{\pi_i \mathcal{N}(\boldsymbol{y}_j;\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)\right\}, \quad (1)$$

$$\log L_{UC}(\boldsymbol{\theta}) = \sum_{j=1}^{n} m_j \log\left(\sum_{i=1}^{K} \pi_i \mathcal{N}(\boldsymbol{y}_j;\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)\right). \quad (2)$$

因此，观测数据的总对数似然为

$$\log L_{PC}^{(\mathrm{ig})}(\boldsymbol{\theta}) = \log L_C(\boldsymbol{\theta}) + \log L_{UC}(\boldsymbol{\theta}). \quad (3)$$

B. 间隔置信度

间隔置信度通过衡量给定输入对应的后验类别概率中最高值与次高值之间的差异，量化模型对其预测的确定性。对于数据点$\boldsymbol{y}_j$，间隔置信度 δ_j 定义为

$$\delta_j = \tau_{(1)j} - \tau_{(2)j},$$

其中，$\tau_{(1)j}$ 和 $\tau_{(2)j}$ 分别表示最大的后验概率和第二大的后验概率。

对于二元分类，间隔置信度可简化为:

$$\delta_j = \left|2\tau_{1j} - 1\right|, \tag{4}$$

其中，τ_{1j} 是 y_j 属于类别 1 的后验概率。间隔置信度接近 1 表示分类结果有较高置信度，而接近 0 则对应较高的预测不确定性。

根据补充材料（S1 节）中的推导，间隔置信度与熵之间的关系可表示如下:

$$H(\tau) \approx \log 2 - \frac{\delta_j^2}{2}, \tag{5}$$

这揭示了熵与平方边际置信度之间的线性关系。

C. 随机缺失机制

按照 [5], 的框架，我们假设缺失是随机缺失（MAR），这意味着缺失仅取决于观测特征向量。在此背景下，缺失机制定义为

$$\begin{aligned}\Pr(M_j = 1 \mid y_j, z_j) &= \Pr(M_j = 1 \mid y_j) \\ &= q(y_j; \boldsymbol{\theta}, \boldsymbol{\xi}),\end{aligned}$$

其中，$M_j \in \{0,1\}$ 表示观测 j的缺失指示变量，$M_j = 1$ 表示标签缺失，$M_j = 0$则表示标签未缺失；$\boldsymbol{\xi} = (\xi_0, \xi_1)$用于参数化缺失机制。

接下来，缺失概率被建模为对数熵 $\log H(y_j;\theta)$ 的逻辑函数。根据补充材料（S2 节）中的推导，$\log H(y_j;\theta)$ 与 δ_j^2 呈近似线性关系。因此，得到一个新的缺失概率模型:

$$q(y_j; \boldsymbol{\theta}, \boldsymbol{\xi}) = \frac{\exp\left(\xi_0 + \xi_1 \delta_j^2\right)}{1 + \exp\left(\xi_0 + \xi_1 \delta_j^2\right)}.$$

D. Aranda–Ordaz 链接函数

基于此，AO 链接函数通过允许协变量对缺失概率产生不对称影响，对 logit 链接函数进行了推广，其定义如下:

$$g(q;\lambda) = \begin{cases} \log\left(\dfrac{(1-q)^{-\lambda} - 1}{\lambda}\right) & \lambda \neq 0, \\ \log\big(-\log(1-q)\big) & \lambda = 0, \end{cases}$$

其中，q 是缺失概率，λ 是控制不对称程度的形状参数。

在MAR假设下，我们用平方边际置信度对缺失概率建模。因此，q可写为 $q(\delta_j;\ \boldsymbol{\alpha}, \lambda)$，其中 $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)$用于参数化缺失机制。相应的逆链接函数形式为

$$\begin{gathered} q(\delta_j; \boldsymbol{\alpha}, \lambda) = g^{-1}(\eta; \boldsymbol{\alpha}, \lambda) = 1 - \left(1 + \lambda e^{\eta}\right)^{-1/\lambda}, \\ \text{where } \eta = \alpha_0 + \alpha_1 \delta_j^2. \end{gathered}$$

在实际应用中，参数 λ 可以预先确定。或者，如果希望获得更高的建模准确率，也可以将 λ 纳入期望条件最大化算法，并进行迭代更新。

我们将完整参数集定义为 $\boldsymbol{\Theta} = \{\lambda, \boldsymbol{\theta}, \boldsymbol{\alpha}\}$。代入上述表达式可得结果为

$$q(\boldsymbol{y}_j; \boldsymbol{\Theta}) = 1 - \left(1 + \lambda e^{\alpha_0 + \alpha_1 \delta_j^2}\right)^{-1/\lambda}. \tag{6}$$

E. 随机缺失机制下的对数似然函数

为将随机缺失机制纳入高斯混合模型的似然中，训练集的密度函数根据缺失指示变量 m_j 划分为 $f(\boldsymbol{y}_j, z_j, m_j = 0;\ \boldsymbol{\Theta})$ 和 $f(\boldsymbol{y}_j, m_j = 1;\ \boldsymbol{\Theta})$。给定 $m_j = 0$,

$$\begin{aligned} &f(\boldsymbol{y}_j, z_j, m_j = 0; \boldsymbol{\Theta}) = \\ &\prod_{i=1}^{K} \Big\{\pi_i \mathcal{N}(\boldsymbol{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\Big\}^{z_{ij}} \Big\{1 - q(\boldsymbol{y}_j; \boldsymbol{\Theta})\Big\}. \end{aligned}$$

给定 $m_j = 1$,

$$f(\boldsymbol{y}_j, m_j = 1; \boldsymbol{\Theta}) = \sum_{i=1}^{K} \pi_i \mathcal{N}(\boldsymbol{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\, q(\boldsymbol{y}_j; \boldsymbol{\Theta}).$$

基于这两个不同的密度函数， $\boldsymbol{\Theta}$ 的完整似然函数为

$$L_{PC}^{(\text{full})}(\boldsymbol{\Theta}) = \prod_{j=1}^{n} \Big\{ f(\boldsymbol{y}_j, z_j, m_j = 0; \boldsymbol{\Theta}) \Big\}^{1-m_j} \Big\{ f(\boldsymbol{y}_j, m_j = 1; \boldsymbol{\Theta}) \Big\}^{m_j}.$$

近期研究表明， $\boldsymbol{\Theta}$的完整对数似然函数可以表示为

$$\log L_{PC}^{(\text{full})}(\boldsymbol{\Theta}) = \log L_{PC}^{(\text{ig})}(\boldsymbol{\theta}) + \log L_{PC}^{(\text{miss})}(\boldsymbol{\Theta}), \quad (7)$$

其中

$$\log L_{PC}^{(\text{miss})}(\boldsymbol{\Theta}) = \sum_{j=1}^{n} \Big[(1 - m_j)\, \log\{1 - q(\boldsymbol{y}_j; \boldsymbol{\Theta})\} + m_j\, \log q(\boldsymbol{y}_j; \boldsymbol{\Theta}) \Big], \quad (8)$$

对数项 $\log\, L_{PC}^{(\text{ig})}(\boldsymbol{\theta})$ 和 $q(\boldsymbol{y}_j; \boldsymbol{\Theta})$ 在方程 3 和 6 [14] 中作了形式化定义。

为了获得$\boldsymbol{\Theta}$的最大似然估计，我们采用 ECM。详细步骤见补充材料（S4 节），下方则给出了示意流程图。

算法 1： 用于最大化 对数 $L_{PC}^{(\text{full})}(\boldsymbol{\Theta})$ 的期望条件最大化算法

输入： 观测数据 $\{\boldsymbol{y}_j, m_j\}_{j=1}^{n}$；初始值 $\boldsymbol{\Theta}^{(0)}$

输出： $\widehat{\boldsymbol{\Theta}}$

对于 $t = 0$、1、2.....直到收敛，**执行**

// E 步

计算 $\tau_{ij}^{(t)}$ 、 $\delta_j^{(t)}$ 、 $q_j^{(t)}$ 。

// CM 步骤 1

更新 $\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q_1(\boldsymbol{\theta})$ ；(拟牛顿法)。

// CM 步骤 2

更新 $(\boldsymbol{\alpha}^{(t+1)},\ \lambda^{(t+1)}) = \arg\max_{\boldsymbol{\alpha}, \lambda} Q_2(\boldsymbol{\alpha}, \lambda)$ ；(牛顿法/其他方法；对 λ 进行一维线搜索)。

III. 近似有效性

图 1 展示了对数熵 $\log(H(m))$ 与平方边际置信度 m^2之间的关系。对于 $m^2 < 0.36$，二阶泰勒近似与真实熵函数十分接近，这表明在低间隔区域二者近似呈线性关系。

基于这一观察，我们采用 m^2 =0.36 作为阈值。对于满足 $m^2 \le 0.36$ 的样本——这对应于 $[0.2, 0.8]$ 中的后验类别概率，且在实际中涵盖了大多数观测值——平方边际置信度在对缺失机制建模时，能够有效且高效地替代香农熵。

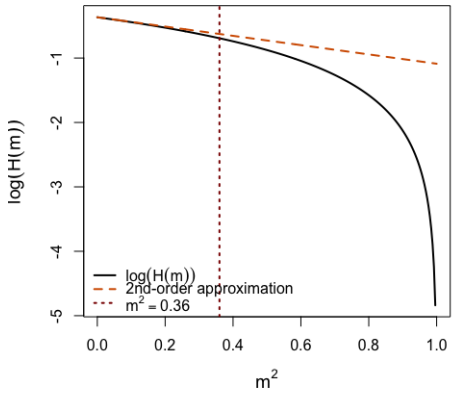

图 1：实线表示精确的香农熵，虚线表示其二阶泰勒近似。竖线表示 $m^2 = 0.36$。

IV. 模拟

A. 模拟可视化

为研究第 II 节所述的随机缺失机制，我们在方差矩阵相同的双分量高斯混合模型下进行了受控模拟。数据生成过程产生了重叠密度的观测值，因此在分类边界周围形成了模糊度较高的区域。数据生成后，

类别标签按照随机缺失机制被删除，并经过校准，使总体缺失率达到 70%。所得数据集如图 2 所示，其中展示了有标签观测和无标签观测的空间分布，以及缺失标签的总体比例。

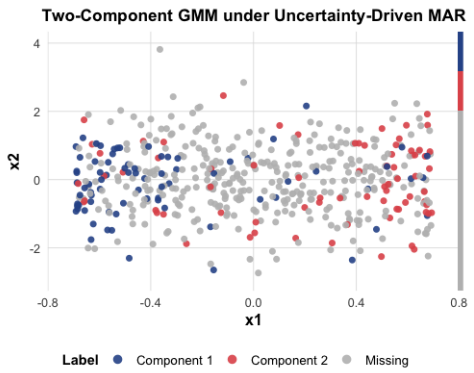

图2：蓝色和红色点表示两个组成部分，灰色点表示缺失标签。右侧条带显示它们的比例。

B. 模拟评估

为评估显式建模随机缺失机制对分类性能的影响，我们对模拟数据集应用了三种估计量。我们从判别能力、概率校准和预测性能三个方面评估了每种方法。定量结果见表 I，相关 ROC 曲线见图3。

结果表明，两种 ECM 变体的表现始终优于逻辑回归基线。它们的曲线下面积略高（0.71 对比 0.705），概率校准显著改善，对数损失（0.67 对比 0.97）和 Brier分数（0.233 对比 0.368）均明显降低。在根据 Youden指数选定的阈值下，两种 ECM 变体的 F1 分数均超过 0.70；logit 链接函数更有利于提高召回率，而 Aranda–Ordaz 链接函数则带来略高的精确率。在更复杂的缺失机制场景中，Aranda–Ordaz形式所具备的额外灵活性可能更具优势（第 IV-C 节）。

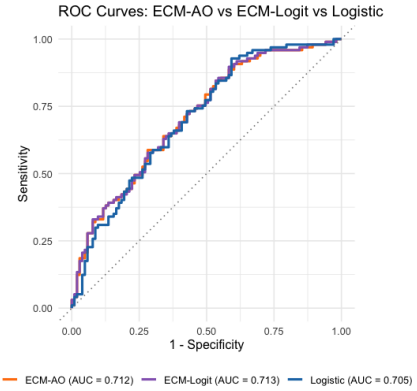

图3：比较随机缺失条件下三种估计量的ROC曲线。

此外，我们比较了 ECM–AO 和逻辑回归在不同缺失标签比例下的表现。图4显示，随着缺失标签比例增加，表现呈现出明显趋势。ECM–AO 的判别能力保持稳定，而逻辑回归的表现则逐渐下降。这种差异反映了随机缺失机制：仅使用观测到的标签进行训练会导致选择偏差，而显式建模缺失情况则使 ECM–AO 能够在有标签信息逐渐减少时，保持判别能力和校准。

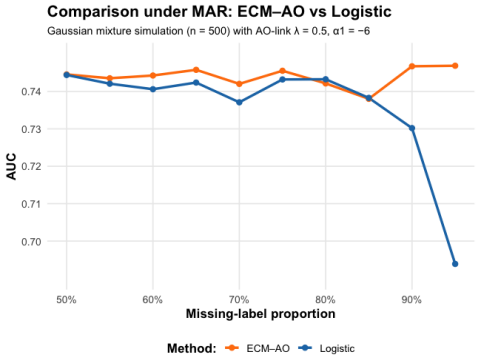

图4：两种估计量在不同缺失比例下的曲线下面积

T表 I：随机缺失条件下三种估计量的性能指标比较。

方法	AUC	对数损失	布里尔分数	精确率$_{opt}$	召回率$_{opt}$	$F1_{opt}$
ECM（AO 链接函数）	0.712	0.675	0.233	0.599	0.845	0.701
ECM（logit 链接函数）	0.713	0.673	0.233	0.592	0.897	0.713
逻辑回归基线	0.705	0.971	0.368	0.143	0.072	0.096

C. 算法鲁棒性验证

为严格评估分布设定错误情况下的稳健性，我们将分析扩展到高斯假设之外，并在三种非高斯混合设置下评估所提出的 ECM–AO估计量：伽马、贝塔和拉普拉斯混合。这些设置共同违背了若干关键的高斯假设，包括对称性、无限支撑集和尾部行为 [25]–[27]。这为稳健性提供了严格的检验。

为实施该测试，我们在三种二维非高斯混合模型场景下评估了 ECM–AO 估计量，并将其与忽略缺失机制的逻辑回归基线进行了比较。如表 II 所示，ECM–AO 在所有场景下的表现均相当或更优。尽管不同混合分布下的曲线下面积值保持稳定，ECM–AO 仍取得了显著更低的对数损失和 Brier分数，并保持了有利的精确率–召回率权衡：召回率更高，且精确率没有明显损失。相比之下，逻辑回归在随机缺失条件下的校准和召回率均有所下降。这些结果表明，所提出的框架在中等程度的分布设定错误下仍具有鲁棒性。

V. MAGIC GAMMA TELESCOPE 数据集

A. 数据描述

本研究的实证分析基于从 UCI机器学习仓库获取的 MAGIC Gamma Telescope 数据集。该数据集包含 19,020 个观测值。每个观测值由十个连续图像变量以及一个二元类别标签描述，该标签表示真实的粒子类型（伽马粒子 或强子）。

经过数据预处理后，保留了四个具有代表性的变量 *fAlpha*、*fLength*、*fM3Long* 和 *fSize*，用于后续建模。图 5 显示，成对散点图矩阵中两类粒子存在明显重叠，这表明简单的线性边界不足以进行判别。

基于上述数据集，我们在 MAR假设下构建了一种受控的缺失机制。该过程导致约 70% 的类别标签缺失，从而生成一个用于实证评估的部分标注的MAGIC数据集。

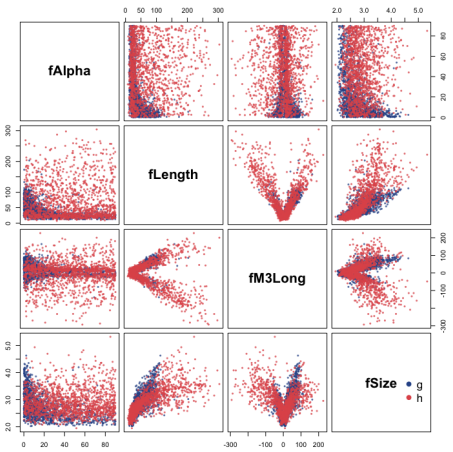

图 5：MAGIC 数据集中四个变量的成对散点图矩阵。蓝色和橙色点分别表示伽马粒子 和强子 事件。

B. 模型比较

为进一步评估 ECM–AO 的分类性能，我们在不同的分类阈值下将其与逻辑回归进行比较。

图 6 展示了分类阈值在 0.3 到 0.7 之间变化时的性能。在此范围内，与逻辑回归相比，ECM-AO 保持了更高的准确率，且召回率和精确率更加稳定。相比之下，

表 I:　　不同生成混合分布下随机缺失的性能指标比较　　s.

分布	方法	AUC	对数损失	布里尔分数	精确率opt	召回率opt
二维伽马混合分布	ECM（AO 链接函数）	0.732	0.862	0.244	0.664	0.778
	逻辑回归基线	0.744	1.173	0.431	0.350	0.485
二维 Beta 混合	ECM（AO 链接函数）	0.668	0.683	0.241	0.609	0.857
	逻辑回归基线	0.689	1.107	0.402	0.257	0.184
二维 Laplace 混合	ECM（AO 链接函数）	0.696	0.677	0.234	0.670	0.702
	逻辑回归基线	0.686	1.006	0.377	0.365	0.337

随着阈值升高，逻辑回归模型的准确率和召回率明显下降，表明其对截断值选择更为敏感。这种稳定性表明，ECM-AO 在不同阈值下能够生成更平滑、更稳健的决策边界。

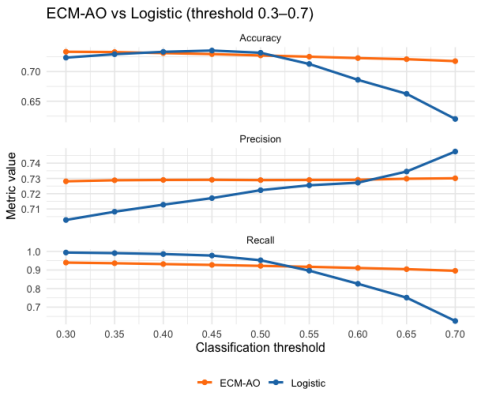

图 6：在 70% 标签缺失情况下，ECM–AO 和逻辑回归在不同分类阈值（0.3–0.7）下的精确率、召回率和准确率。

C. 缺失比例对模型性能的影响

如图 7 所示，我们比较了在缺失比例逐渐增加（50–90%）时 ECM–AO 与逻辑回归的表现。对于每个缺失水平，我们都在随机缺失机制下生成多个数据集，并以重复运行所得曲线下面积的均值汇总性能。正如预期，随着缺失率增加，两种模型的分类性能都在下降。在中等缺失水平（50%–70%）下，ECM–AO 的表现始终优于逻辑回归基线。然而，随着缺失程度增加，性能差距逐渐缩小，并在缺失率接近 90% 时反转。这种现象与极端标签稀疏性下基于随机缺失的非确定性建模所含信息减少相符；在这种情况下，自适应 AO 链接函数的优势不再那么明显。

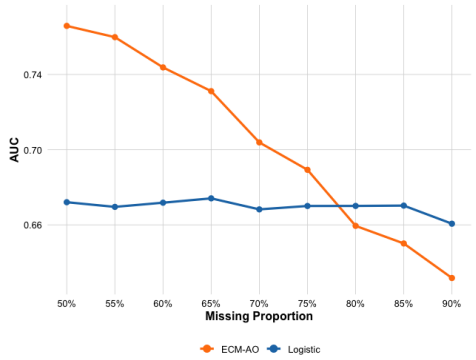

图 7：不同缺失比例下 ECM–AO 与逻辑回归的曲线下面积比较。点表示多次实现的平均曲线下面积。

VI. 结论

本研究表明，在半监督框架中显式建模标签缺失，能够在随机缺失（MAR）条件下实现更稳健、校准效果更好的分类。在模拟数据和真实数据上，所提出的 ECM–AO 始终优于忽略缺失机制的逻辑回归基线。它在改善校准效果的同时，在

中等程度的标签缺失下保持了稳定的判别能力。AO 链接函数所带来的额外灵活性，使不确定性能够更平滑地映射为缺失概率，从而得到更可靠的后验估计。尽管在极端缺失情况下其优势有所减弱，ECM–AO 仍保持数值稳定，且具有竞争力。

尚待解决的一个挑战是将该框架扩展到包含两个以上成分的混合模型，因为当前的间隔置信度依赖于两个最大的后验概率。开发能够利用完整后验向量的不确定性度量，是扩展到更复杂混合模型场景的一个有前景方向。

VII. 补充材料

补充材料中提供了进一步的理论推导和算法细节（第 S1–S5 节），可通过以下网址获取：

https://github.com/LeomusUNSW/IJCNN

参考文献

[1] J. E. Van Engelen 和 H. H. Hoos，“半监督学习综述”，*Machine Learning*，第 109 卷，第 2 期，第 373–440 页，2020年。[2] X. Zhu，“半监督学习文献综述”，威斯康星大学麦迪逊分校计算机科学系，技术报告，2005年。[3] H. Wu 和 S. Prasad，“利用伪标签进行高光谱图像分类的半监督深度学习”，*IEEE Transactions on Image Processing*，第 27 卷，第 3 期，第 1259–1270 页，2017年。[4] A. Martinez-Uso、F. Pla 和 J. M. Sotoca，“用于图像分割的半监督高斯混合模型”，载于*2010 20th International Conference on Pattern Recognition*，2010年，第 2941–2944 页。[5] D. B. Rubin，“推断与缺失数据”，*Biometrika*，第 63 卷，第 3 期，第 581–592 页，1976年。[6] R. J. Little，“针对含缺失值多变量数据的完全随机缺失检验”，*Journal of the American statistical Association*，第 83 卷，第 404 期，第 1198–1202 页，1988年。[7] D. F. Heitjan 和 S. Basu，“区分‘随机缺失’与‘完全随机缺失’”，*The American Statistician*，第 50 卷，第 3 期，第 207–213 页，1996年。[8] D. A. Bennett，“在我的研究中如何处理缺失数据？”*Australian and New Zealand journal of public health*，第 25 卷，第 5 期，第 464–469 页，2001年。[9] X. Yan、W. Xiong、L. Hu 和 F. Wang，“基于高斯混合模型的物联网缺失值填补”，*Mathematical Problems in Engineering*，2015年。[10] F. Mealli 和 D. B. Rubin，“阐明随机缺失及相关定义，以及其与可交换性结合时的含义”，*Biometrika*，第 102 卷，第 4 期，第 995–1000 页，2015年。[11] D. Ahfock 和 G. J. McLachlan，“一个表面悖论：基于部分分类样本的分类器，其期望错误率可能低于样本完全分类时的期望错误率”，*Statistics and Computing*，第 30 卷，第 6 期，第 1779–1790 页，2020年。[12] S. Nakagawa，“缺失数据：机制、方法与启示”，载于*Ecological Statistics: Contemporary Theory and Application*，D. Gibson 和 J. Newman 编。牛津大学出版社，2015年，第 81–105 页。[13] M. S. Santos、R. C. Pereira、A. F. Costa、J. P. Soares、J. Santos 和 P. H. Abreu，“合成缺失数据的生成：按缺失机制进行的综述”，*IEEE Access*，第 7 卷，第 11 651–11 667 页，2019年。[14] Z. Lyu、D. Ahfock、R. Thompson 和 G. J. McLachlan，“在 R 中结合缺失数据机制的半监督高斯混合建模”，*Australian & New Zealand Journal of Statistics*，第 66 卷，第 2 期，第 146–162 页，2024年。[15] Z. Lyu，“在指定缺失数据机制下高斯混合模型贝叶斯规则估计的分析”，*Computational Statistics*，第 39 卷，3751，2024 7，3727 第。—。[16] G. J. McLachlan 和 D. Peel，有限混合模型。约翰 · 威立父子出版公司，2000 年。[17] G. J. McLachlan、S. X. Lee 和 S. I. Rathnayake，“有限混合模型”，*Annual review of statistics and its application*，第 6 卷，第 1 期，第 355–378 页，2019年。[18] S. Zhang，“KNN-CF 方法：将确定性因子融入 KNN 分类”，*IEEE Intell. Informatics Bull.*，第 11 卷，第 1 期，第 24–33 页，2010年。[19] J. G. Ibrahim、S. R. Lipsitz 和 M.-H. Chen，“缺失数据机制为非可忽略时广义线性模型中的缺失协变量”，*Journal of the RoyalStatistical Society: Series B (Statistical Methodology)*，第 61 卷，第 1 期，第 173–190 页，1999年。[20] W. Wang 和 M.-L. Zhang，“通过置信度评分间隔最大化实现半监督部分标签学习”，载于神经信息处理系统进展，第 33 卷，2020年，第 6982–6993 页。[21] F. J. Aranda-Ordaz，“二元响应数据加性变换的两个变换族”，*Biometrika*，第 68 卷，第 2 期，第 357–363 页，1981年。[22] V. Scalera、M. Iannario 和 A. C. Monti，“稳健链接函数”，*Statistics*，第 55 卷，第 4 期，第 963–977 页，2021年。[23] G. Pumi、C. Rauber 和 F. M. Bayer，“采用 Aranda–Ordaz 链接函数的 Kumaraswamy 回归模型”，*TEST*，第 29 卷，第 4 期，第 1051–1071 页，2020年。[24] H. M. Dehbi、M. Cortina-Borja 和 M. Geraci，“用于学生表现评估的 Aranda–Ordaz 分位数回归”，*Journal of Applied Statistics*，第 43 卷，第 1 期，第 58–71 页，2016年。[25] A. R. Webb，“用于目标识别的伽马混合模型”，*Pattern Recognition*，第 33 卷，2054，2000 12，2045 第。—。[26] A. K. Gupta 和 S. Nadarajah，贝塔分布及其应用手册。CRC出版社，2004年。[27] W. Song、W. Yao 和 Y. Xing，“基于拉普拉斯分布的稳健混合回归模型拟合”，*Computational Statistics& DataAnalysis*，第 71 卷，第 128–137 页，2014年。