IEEE 版权声明

本作品已提交给 IEEE，供其考虑出版。版权可能会在未通知的情况下转让，此后本版本可能无法再访问。

arXiv:2304.00537v1 [stat.ME] 2 Apr 2023

面向多变量零膨胀连续数据的基于 Copula 的密度估计模型

Keita Hamamoto[1*]

[1]株式会社日立制作所中央研究实验室，国分寺 185-8601，日本 n

摘要

零膨胀连续数据普遍存在于许多领域，其中观测到大量严格为零的零值数据，而其他数据则呈连续分布。由于其分布中离散性与连续性的混合结构，统计分析具有挑战性，尤其是在多变量情形下。本文提出了两种基于 Copula 的密度估计模型，用于处理零膨胀连续变量之间的多变量相关性。为克服零膨胀数据中并列数据问题导致的 Copula 使用困难，我们提出了一种新型 Copula——修正高斯 Copula，并给出了高效的参数估计和似然计算方法。数值实验表明，与传统密度估计方法相比，本文提出的方法具有优越性。

1 引言

密度估计是统计机器学习、异常检测等许多应用中最重要的任务之一。密度估计的目标是根据观测数据估计概率密度函数（PDF）。估计得到的 PDF 能够提供有关数据属性的丰富信息，包括偏度、多峰性、聚类结构、相关性等，可直接用作后续分析和应用的构件。针对多元连续数据，已有有效的密度估计工具，例如多元高斯模型、高斯混合模型（GMM）和核密度估计（KDE） [1, 2]。多元离散分布也有相应的处理方法，例如一系列多元泊松分布 [3], ，其中会对联合概率质量函数进行估计。

对于混合随机变量，即连续变量与离散变量的混合，其分布建模需要特别谨慎，参见 [4]。零膨胀连续数据属于混合随机变量，它同时包含连续观测值和大量恰好为零的观测值。对于零膨胀离散变量，已有广为人知的参数模型来描述其分布，例如零膨胀泊松分布 [5] 和零膨胀负二项分布 [6], ；而零膨胀有时也会出现在连续变量中。生物医学领域中的药物剂量、酒精摄入量和某种生物标志物的含量 [7], ，以及金融领域中的未偿贷款金额、配偶收入和抵押品价值，都是零膨胀连续数据的典型例子。对于许多现实应用而言，这类零膨胀现象不可避免，因为它们会自然出现，例如从多个来源收集数据时，如管理组／对照组，或贷款服务中的单身人士／已婚人士。

在一维情况下，零膨胀连续变量的概率密度函数形式为

$$f(x) = q\delta(x) + (1-q)\tilde{f}(x) \tag{1}$$

其中， q 表示零值出现的概率， $\tilde{f}(x)$是非零数据的概率密度函数。狄拉克δ函数 $\delta(x)$ 在原点处发散，在其他位置为零，用于表示点质量 [4] 的无限密度。本文将重点研究零膨胀非负连续数据，其中$\tilde{}$，且 $\tilde{f}(x)$仅在 $x > 0$ 时取有限值。原点处概率密度函数发散、正值域内密度有限的点质量具有不同的概率测度，常会形成多模态结构。在多元情形下，概率密度函数的建模要困难得多。对于 D元零膨胀数据，每个数据点都位于 2^D 个子空间之一。每个子空间对应于 D 个变量中一组取零值的变量，且维数各不相同。例如，对于 $D = 3$，联合概率是以下部分的混合：原点处的一个点质量、沿各坐标轴的三条线密度、由两条坐标轴张成的三个平面上的面密度，以及一个体密度。本文将密度一词用于指代上述所有类型的密度，有时也包括概率质量。

*keita.hamamoto.uv@hitachi.com

这些发散的概率密度函数、呈指数级增长的子空间数量以及不同的维数，使多元零膨胀连续数据的密度估计难以实现。一种可能的方法是采用蛮力法，逐个子空间进行密度估计。然而，由于高维情况下子空间数量呈指数级增长，计算成本会变得难以承受。此外，随着维数升高，位于每个子空间中的数据点数量呈指数级减少，导致逐个子空间进行密度估计的性能下降。传统高斯混合模型能够处理多峰分布，但除非使用呈指数级增长数量的高斯分量，否则也无法捕捉所有子空间。此外，即使在低维情形下，参数估计也常会导致沿坐标轴出现无限窄且陡峭的高斯分布，从而产生不合理地偏大的似然值，而对数据的拟合效果却很差。核密度估计同样无法捕捉不同的维数。由于密度核的适当带宽应高度依赖其所在子空间的维数，估计得到的概率密度函数会沿垂直于各子空间的方向过于宽泛，并在数据稀疏区域赋予较大的概率密度函数值。此外，众所周知，核密度估计在高维情况下表现不佳，这是由于所谓的“维数灾难 [8]。因此，我们需要一种能够妥善处理多元零膨胀连续数据的新型密度估计方法。

本文提出两种基于 Copula 的密度估计方法，分别处理不同的零膨胀机制。我们采用式（1）中的混合随机变量模型来表示边缘分布，并使用 Copula 密度捕获多元相关性。这种多元混合随机变量建模使我们能够处理不同维数的子空间，例如，概率质量、线密度、面密度和体密度都可以用统一的方式表示，这与传统的高斯混合模型和核密度估计不同。在第二个模型中，我们开发了一种新型 Copula——修正高斯 Copula，以避免将 Copula 应用于零膨胀数据时与并列数据相关的困难。我们还展示了用于估计模型参数和计算似然值的多项式时间算法。针对合成数据和真实数据的数值实验表明，与传统的高斯混合模型和核密度估计相比，我们的方法性能更高。

2 预备知识

本节将介绍一些基本概念和符号，以便更好地理解我们提出的方法。首先，我们简要概述 Copula 的符号和用法。我们将说明，直接将传统 Copula 应用于零膨胀数据并不恰当，这是因为存在众所周知的并列数据问题。最后，我们将结合缺失数据分析中提出的概念，介绍零膨胀的两种不同机制。

2.1 Copula

Copula 是统计学中用于多元建模的一种范式，凭借其强大的能力和灵活性，日益受到关注 [9]。自 Sklar 首次提出 Copula 并经历后续一系列发展以来 [10–12], ，它已应用于金融 [13, 14], 、可靠性分析 [15],、生存分析 [16], 、临床医学 [17], 等多个领域。采用 Copula 的表述中，多元概率密度函数写作

$$f(x_1,\ldots,x_D)=\left[\prod_{i=1}^{D} f_i(x_i)\right] c(F_1(x_1),\ldots,F_D(x_D)). \quad (2)$$

f_i 是变量 x_i 的边缘概率密度函数。最后一个因子c是边缘累积分布函数（CDF） $F_i(x_i)$ 的多元函数，表示 Copula 密度，用于描述特征之间的相关性。当变量相互独立时，Copula 密度为 1；但一般情况下并非如此。对于这种形式的概率密度函数，可以分别选择合适的方法来建模边缘分布和 Copula 密度。一种典型的选择是半参数 Copula，即使用灵活的非参数模型来描述边缘分布，并使用参数模型来描述 Copula 密度。高斯 Copula 是参数 Copula 族中的一个简单成员，它基于多元正态分布的简单结构，能够处理高维（$D \geq 3$）相关性。高斯 Copula 的 Copula 密度定义为

$$c(F_1(x_1),\ldots,F_D(x_D)) = \left.\frac{\phi_D(\omega_1,\ldots,\omega_D|\mathbf{0},\Sigma)}{\prod_{i=1}^{D}\phi_1(\omega_i)}\right|_{\omega_i=\Phi^{-1}\circ F_i(x_i)} \quad (3)$$

其中， ϕ_1 和 Φ 分别是单维标准正态分布的概率密度函数和累积分布函数。ϕ_D是均值向量为零、协方差矩阵为Σ的 D 维多元正态分布的概率密度函数，而 F_i 是 x_i的累积分布函数。对概率变量$x_i \to \omega_i = \Phi^{-1} \circ F_i(x_i)$进行非线性单调非递减变换，以消除边缘分布信息，并使所有变量都服从标准正态变量$\omega_i \sim \mathcal{N}(0,1)$。当分布 F_i 在其

支撑集上连续时，这种向正态变量的变换是可行的。经过这些非线性变换后，变量间的相关性通过协方差为Σ的多元正态分布来描述。ϕ_D 的均值向量为零，且由于所有边缘分布均为标准正态分布， Σ 的对角元素为 1。式 (3) 分母中的因子是为了使概率密度函数的积分为 1。由于概率测度在变换下按 $f_i(x_i)\mathrm{d}x_i = \phi_1(\omega_i)\mathrm{d}\omega_i$ 改变，因此通过 $\int \mathrm{d}x_1 \ldots \mathrm{d}x_D f = \int \mathrm{d}\omega_1 \ldots \mathrm{d}\omega_D \phi_D = 1$确保归一化。通常在 ω空间中使用最大似然估计 (MLE) 来估计 Σ，其中观测到的 ω的经验协方差矩阵用作 Σ的估计量。

Copula 使用中一个广为人知的问题是如何处理并列数据 [18,19]。如果许多数据点取值相同，变换 $\Phi^{-1} \circ F_i$ 就会变得不连续，在并列点处出现跳跃，因此变换后的值ω_i 不再服从 $\mathcal{N}(0, 1)$。由于零膨胀是并列数据的一种，直接应用 Copula 可能导致参数估计有偏且拟合效果不佳。在我们的第二个模型中，我们展示了变换 $\Phi^{-1} \circ F_i$的不连续性，说明了ω_i适用的边缘分布，并开发了一种处理并列数据的新型 Copula。

2.2 零膨胀的机制

为了更好地理解零膨胀数据的统计特性，需要明确其产生机制。这里我们展示两种零膨胀机制，并将其与缺失数据分析中的概念 [20–22]联系起来。

在第一种情形中，零膨胀仅仅意味着信息缺失，即假设数据中的零值分量是完全随机获得的。我们假设首先有 D 维正值数据，然后对其应用 D 维二值掩码，最终得到 D 维零膨胀数据。该掩码 D维相关伯努利型分布，例如可以从受限玻尔兹曼机（RBM）[23, 24], 中抽取；它与原始正值数据不相关。参照缺失数据分析中的类似概念“完全随机缺失”（MCAR） [22]，我们将这种情形称为“完全随机零膨胀”（ZICAR）。当 MCAR 类型的缺失值以零值填补时，就可能出现这种零膨胀情形。

另一种情形则相当自然。当我们遇到零值数据时，有时可以理解为实际存在一个远小于某个阈值的值。例如，该阈值可能由测量设备的分辨率或数据累积过程中的舍入规则决定。在这种情况下，完全随机缺失类型的缺失情形并不适用，因为零值出现的概率强烈依赖于原本应获得的值。这种情况是随机缺失（MAR）的一种特殊情形 [22] ，其中零值出现是确定性的。我们将这种情形称为阈值化零膨胀（ZIBT）。

在后续章节中，我们分别给出适用于 ZICAR 和 ZIBT 类型的两种密度估计模型。

3 提案 1：ZICAR 情形的密度估计

如前所述，多变量零膨胀数据的密度估计难点在于子空间数量呈指数级增长，且子空间的维数各不相同。由于逐子空间估计的复杂度呈指数级增长，而且每个子空间中的数据点数量较少，估计结果并不可靠，因此我们需要设法尽可能利用更多数据点来估计一个多峰的 D 变量密度。

本节和下一节提出的两种模型，其基本策略是设计密度模型，使所有在 D 维中定义的模型参数都能在低维空间中估计。正如后文将展示的，基于 copula 的密度模型具有可边缘化性，因此 D 维密度中的参数会直接出现在低维边缘化密度中，从而我们可以在低维空间中估计高维参数的子集。当边缘化至低维，即二维时， 2^D 子空间会投影到仅有 $2^2 = 4$ 个子空间上，这样我们便可以充分利用几乎所有数据点来估计参数。

3.1 密度模型

在我们的第一个模型中，针对 ZICAR 类型情形进行专门设计，我们假设正值的 D 维父分布，其概率密度函数记为 $g_D(x_1, \ldots, x_D)$，其中不包含零膨胀。假设含零膨胀的观测数据由对 g_D 应用 D 维二值掩码生成，如 ZICAR 的定义所述。我们使用高斯 Copula 模型将 g_D 分解为边缘分布和 Copula 密度，如下所示

$$g_D(x_1, \ldots, x_D) = \left[\prod_{i=1}^{D} g_{\{i\}}(x_i)\right] \left.\frac{\phi_D(\omega_1, \ldots, \omega_D | \mathbf{0}, \Sigma)}{\prod_{i=1}^{D} \phi_1(\omega_i)}\right|_{\omega_i = \Phi^{-1} \circ G_i(x_i)} \tag{4}$$

其中 $g_{\{i\}}(x_i)$ 为边缘概率密度函数，G_i 为其累积分布函数。与多元正态分布类似，高斯 Copula 模型具有可边缘化性，这是我们的方法中反复利用的一项关键特性。借助变量替换公式 $f_i(x_i)\mathrm{d}x_i = \phi_1(\omega_i)\mathrm{d}\omega_i$ 以及 ϕ_D 的可边缘化性，可以轻松地将概率密度函数 g_D 边缘化到任意变量子集 ($S \subseteq D$)，如下所示

$$\begin{aligned} g_S(\boldsymbol{x}_S) &= \int \mathrm{d}\boldsymbol{x}_{\bar{S}}\; g_D(x_1, \ldots, x_D) \\ &= \left[\prod_{i\in S} g_{\{i\}}(x_i)\right] \left.\frac{\phi_{|S|}(\boldsymbol{\omega}_S|\mathbf{0}, \Sigma_S)}{\prod_{i\in S}\phi_1(\omega_i)}\right|_{\omega_i = \Phi^{-1}\circ G_i(x_i)} \end{aligned} \tag{5}$$

$\bar{}$与 $S \subseteq \{1, \ldots, D\}$、$S = \{1, \ldots, D\}\backslash S$、$\boldsymbol{x}_S = \{x_i | i \in S\}$和 $\int \mathrm{d}\boldsymbol{x}_{\bar{S}} = \prod_{i\in\bar{S}} \int_0^\infty \mathrm{d}x_i$。本文将积分测度写在被积函数前，以明确积分范围。协方差参数 $\Sigma_S = \{\Sigma_{ij}\}_{i\in S,\ j\in S}$ 是原始Σ 的方形子矩阵，仅恢复 S 中的维度。为简化记号，有时将单元素 $\{i\}$写作 i，并将完整特征集 $\{1, \ldots, D\}$ 简写为 D。现在，我们可以提出第一个模型，用来表示掩码处理后零膨胀数据的总联合概率密度函数

$$f_D(x_1, \ldots, x_D) = \sum_{S\subseteq D} q_S \delta(\boldsymbol{x}_{\bar{S}}) g_S(\boldsymbol{x}_S) \tag{6}$$

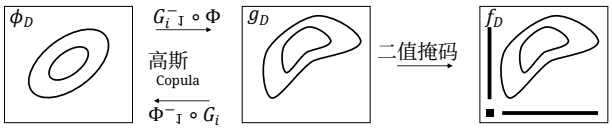
其中，$\delta(\boldsymbol{x}_{\bar{S}}) \equiv \prod_{i\in\bar{S}} \delta(x_i)$定义了 $|S|$ 维子空间。$q_S \equiv P(\boldsymbol{x}_{\bar{S}} = 0, \boldsymbol{x}_S > 0)$是处理零值出现相关性的掩码分布。对于最简单的情形，即零值出现彼此独立，掩码服从伯努利分布，$q_S = \left[\prod_{i\in\bar{S}} q_i\right]\left[\prod_{j\in S}(1-q_j)\right]$，其边缘零值出现概率为 q_i。

我们第一个模型的整体生成过程如图 1(a) 所示。其概率密度函数为 g_D 的母分布是高斯 copula 模型（中间图），由逐维非线性变换作用于多元正态分布（左图）生成。随后，对 g_D 应用多元二值掩码，生成总概率密度函数 f_D （右图）。

3.2 参数估计

根据式（6），需要从观测到的零膨胀数据中估计的参数包括掩码分布 q_S、边缘分布 $g_{\{i\}}$，以及高斯 Copula 中的协方差矩阵 Σ。要估计掩码分布，可以先通过指示函数 $I(x > 0)$ 将训练数据转换为二元矩阵，其中正值被替换为

(a) ZICAR 的生成模型



(b) 生成用于 ZIBT 的模型

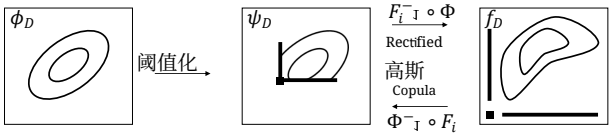


图1：从生成模型的角度总结了我们提出的两种方案。图中以示意方式描绘了分布的等高线。圆点和直线表示零值数据。(a) ZICAR 模型。使用高斯 Copula 表示母分布，其概率密度函数为 g_D。假设多元零膨胀分布由对 g_D 进行 D 元二元掩码生成。(b) ZIBT 模型。本文新提出的修正高斯 Copula 直接对观测到的零膨胀分布进行建模。修正高斯分布及其概率密度函数 ψ_D 是修正高斯 Copula 的关键构件，它由对多元正态分布进行阈值化生成。在该模型中，零膨胀源于阈值化。

1。然后可以使用受限玻尔兹曼机（RBM）等传统估计方法。

将式 (6) 中的总概率密度函数边缘化为单变量分布后，母分布边缘分布的估计便显而易见。利用式 (5) 所示的可边缘化性以及狄拉克 delta 函数的性质，可以对 $\boldsymbol{x}_{D\backslash\{i\}}$ 进行积分，得到

$$\begin{aligned} f_{\{i\}}(x_i) &= \sum_{S\subseteq D} q_S \int \mathrm{d}\boldsymbol{x}_{D\backslash\{i\}} \delta(\boldsymbol{x}_{\bar{S}}) g_S(\boldsymbol{x}_S) \\ &= \sum_{S\subseteq D, i\notin S} q_S \delta(x_i) + \sum_{S\subseteq D, i\in S} q_S g_{\{i\}}(x_i) \\ &= q_{\{i\}}^{\{i\}} \delta(x_i) + q_\emptyset^{\{i\}} g_{\{i\}}(x_i) \end{aligned} \tag{7}$$

其中 $q_{\{i\}}^{\{i\}} = P(x_i = 0) = \sum_{S\subseteq D\backslash\{i\}} q_S$ 和 $q_\emptyset^{\{i\}} = 1 - q_{\{i\}}^{\{i\}} = P(x_i > 0) = \sum_{S\subseteq D\backslash\{i\}} q_{S\cup\{i\}}$ 为边缘掩码分布。由于该表达式与公式 (1) 完全平行，因此可以使用传统方法（如核密度估计），从训练数据中估计母分布$g_{\{i\}}$ 的边缘分布，其中 $x_i > 0$。

高斯 cop- ula Σ 的协方差矩阵估计略为复杂。最简单的方法是，利用训练数据中的 $x_i > 0$ 估计 Σ，适用于所有 $i \in D$

使用变换$\Phi^{-1}\circ G_i(x_i)$后的经验协方差矩阵。然而，由于高维情况下所有元素均为正的数据点数量呈指数级减少，估计会变得不稳定。这里我们展示一种估计 Σ 的方法，该方法充分利用了几乎所有训练数据。与式（7）类似，式（6）中的总概率密度函数可以边缘化为二元形式:

$$f_{\{i,j\}}(x_i,x_j)=q_{\{i,j\}}^{\{i,j\}}\delta(x_i)\delta(x_j)+q_{\{i\}}^{\{i,j\}}\delta(x_i)g_{\{j\}}(x_j) +q_{\{j\}}^{\{i,j\}}\delta(x_j)g_{\{i\}}(x_i)+q_{\emptyset}^{\{i,j\}}g_{\{i,j\}}(x_i,x_j) \quad (8)$$

其中 $q_S^{\{i,j\}}=P(\boldsymbol{x}_S=0,\boldsymbol{x}_{\{i,j\}\setminus S}>0)$ 对应 $S\subseteq\{i,j\}$。在这种形式下，只有最后一项包含 Σ 的信息，更具体地说是 $\Sigma_{i,\,j}$。因此，$\Sigma_{i,\,j}$ 的最大似然估计是 ω_i 与 ω_j 之间的经验协方差，其中仅使用同时满足 $x_i>0$ 和 $x_j>0$ 的数据点来计算协方差。值得注意的是，在完整的 D 维模型中，一个参数，即 $\Sigma_{i,\,j}$，直接出现在二维边缘分布中，因此可以根据二维表达式进行估计。由于可以利用一个象限中的许多数据点进行估计，采用此方法有望提高估计的稳定性。对所有成对的 $1\le i<j\le D$ 重复进行 $\Sigma_{i,\,j}$ 的估计，即可估计出 Σ 的所有分量。

3.3 似然的计算

为了将估计结果用于无监督异常检测等其他应用，需要计算似然函数。然而，由于我们所得到的式（6）中的总概率密度函数包含狄拉克δ函数，概率密度函数的值会发散。本小节将介绍一种处理概率密度函数发散并计算有限似然值的常见方法。

这种概率密度函数发散的情况总是出现在混合随机变量中。例如，在Tobit模型 [25] 和生存分析领域的删失数据 [26], 中，对于混合随机变量中的离散点，其似然应为概率质量。因此，在我们的情形下，与式（6）中的概率密度函数相对应的似然值可写为

$$L_D(x_1,\ldots,x_D)=\sum_{S\subseteq D}q_S I(\boldsymbol{x}_{\bar{S}}=0,\boldsymbol{x}_S>0)g_S(\boldsymbol{x}_S), \quad (9)$$

该式对任何数据都不会产生发散值。特别是，对于满足 $x_{\bar{S}}=0$、$x_S>0$的数据，似然计算为

$$L_D(x_1,\ldots,x_D,\boldsymbol{x}_{\bar{S}}=0,\boldsymbol{x}_S>0)=q_S g_S(\boldsymbol{x}_S). \quad (10)$$

在我们的第二个模型中，我们将再次用指示函数替代狄拉克 delta 函数；在该模型中，我们将使用类似构造的似然函数来表示参数的最大似然估计。

3.4 变量的重新缩放

由于零膨胀连续数据的概率密度函数在每个子空间中的维数不同，似然函数轮廓取决于输入变量的缩放方式。即使在一维情况下，这一点也很明显。由式（1）的概率密度函数构造的似然函数为

$$L(x)=qI(x=0)+(1-q)I(x>0)\tilde{f}(x). \quad (11)$$

由于两个子空间的维数不同，不仅似然值本身会受变量 x 尺度参数的影响，两项之间的平衡也会受到影响。对尺度参数的这种不同依赖是混合随机变量的固有特性；当我们的模型用于例如无监督异常检测时，也会受到这一问题的影响。在此，我们提出一种启发式方法，以完全消除由变量尺度参数引起的问题。在该方法中，我们重新定义变量的尺度参数，使缩放后的似然值满足所需性质。首先，我们使用原始尺度的数据来估计上述边缘分布，并计算边际似然。对于具有 $x_i=0$ 的数据，边际似然的对数值为 $\log q_{\{i\}}^{\{i\}}$；对于 $x_i>0$，则为 $\log\left(1-q_{\{i\}}^{\{i\}}\right)+\log g_{\{i\}}(x_i)$，其中 $q_{\{i\}}^{\{i\}}$ 是估计的边缘零值发生率。我们要求具有$x_i>0$的数据的平均边际似然对数值为 $\log(1-q_{\{i\}}^{\{i\}})$。因此，变量 x_i 应按 $x_i\to x_i/b_i$ 重新缩放，其方式为

$$b_i=\exp\left[-\frac{1}{N_i}\sum_{k,x_i^k>0}\log g_{\{i\}}(x_i^k)\right] \quad (12)$$

其中，N_i 是具有 $x_i^k>0$ 的训练数据数量，x_i^k是第 k个数据的第 i个分量。通过这种重新缩放技巧，我们再次估计边缘分布，从而获得尺度适当的最终估计结果，并使边缘对数似然函数具有所需性质。这种重新缩放方法也适用于我们在下一节提出的第二个模型。

4 提案 2：ZIBT 情形的密度估计

在我们的第一个方案中，假设零膨胀以完全随机方式发生，此时观测到的零仅表示缺少信息，且与

观测到的正值独立。这一假设有时可能不太自然，因为观测到的零通常表示数值很小，低于某个阈值。这里，我们提出第二种密度模型，它假设零膨胀属于 ZIBT 类型。

我们第二个模型所假设的生成过程如图1(b)所示。在该模型中，多元零膨胀数据（右图）直接由 Copula 建模。将高斯 Copula 应用于零膨胀数据，自然需要使用整流高斯分布（中图）；该分布由对多元正态分布（左图）进行阈值化得到，详细说明见后续小节。

4.1 整流高斯分布

在该模型中，我们的最终联合概率密度函数直接通过 Copula 分解，其形式与式(2)类似。然而，由于非线性变换在零膨胀数据上变为不连续，所得变量 $\omega_i = \Phi^{-1} \circ F_i(x_i)$ 不再服从标准正态分布，因此应用高斯 Copula 会导致数据拟合效果不佳。为了说明如何在这种情况下应用 Copula，我们首先展示零膨胀变量的边缘分布和联合分布 ω_i 。

累积分布函数 $F_i(x_i)$ 的正式定义为 $P(X_i \leq x_i)$，其中 X_i 是与其观测值 x_i相关的随机变量。随后，许多带有$x_i = 0$ 的零值数据被映射到同一个位置 $\Phi^{-1} \circ F_i(0) = \Phi^{-1}(q_i)$，其中 $q_i = F_i(0) = P(X_i \leq 0) = P(X_i = 0)$ 是 X_i 中零值出现的概率。另一方面，由于无穷小负值点 -0会被映射到$\Phi^{-1} \circ F_i(-0) = \Phi^{-1}(P(X_i \leq -0)) = \Phi^{-1}(0) = -\infty$，因此变换 $\Phi^{-1} \circ F_i$ 在零点处不连续。由于这一不连续性，变换后的变量 ω_i 不再服从 $\mathcal{N}(0,1)$。对于正值数据 $x_i > 0$，变换为 $\omega_i = \Phi^{-1}\left(q_i + (1-q_i)\,\tilde{F}_i(x_i)\right)$，其中$\tilde{F}_i(x_i) = \int_{-\infty}^{x_i} \mathrm{d}x \tilde{f}_i(x)$ 是正值数据分布的边缘累积分布函数。于是，变换后变量 ω_i的概率密度为 $\left|\frac{\mathrm{d}x_i}{\mathrm{d}\omega_i}\right|(1-q_i)\,\tilde{f}_i = \phi_1(\omega_i)$，因此，正值数据被映射为标准正态变量，仿佛不存在零膨胀一样。 ω_i 的边缘概率密度函数则为，而非 $\phi_1(\omega_i)$，

$$\psi_1(\omega_i|a_i) = q_i\delta(\omega_i - a_i) + I(\omega_i > a_i)\phi_1(\omega_i) \qquad (13)$$

其中 $a_i = \Phi^{-1}(q_i)$ 是分布的阈值参数。等价地， a_i 由条件$q_i = \int_{-\infty}^{a_i} \mathrm{d}\omega\ \phi_1(\omega)$确定。尽管概率密度函数 ψ_1 与标准正态分布的概率密度函数 ϕ_1相似，但不同之处在于，阈值 a_i 以下的 ϕ_1 密度分数被积分并集中在 a_i 处，从而形成概率质量，并用狄拉克 delta 函数表示。由阈值化产生的此类分布称为左删失分布 [26]。更具体地说，对正态变量进行阈值化所生成的分布称为截断高斯分布（RGD） [27, 28] ，并已得到深入研究，例如在因子分析领域 [29]。

本文将 ω_i 的分布称为标准截断高斯分布，因为高斯分量的位置参数和尺度参数分别为 0 和 1。请注意，该分布的均值和标准差不同于位置参数和尺度参数。标准截断高斯分布最简单的生成模型可写为 $\omega_i = \max\{a_i, \nu_i\}$ ，其中 $\nu_i \sim \mathcal{N}$服从 (0, 1)。这里我们引入整流算子

$$\hat{R}_i(a_i) = \delta(\omega_i - a_i)\int_{-\infty}^{a_i} \mathrm{d}\nu_i + \int_{a_i}^{\infty} \mathrm{d}\nu_i\delta(\omega_i - \nu_i) \qquad (14)$$

其中，积分作用于算子右侧的函数。边缘概率密度函数 (13) 随后可简化为

$$\psi_1(\omega_i|a_i) = \hat{R}_i(a_i)\phi_1(\nu_i). \qquad (15)$$

将标准截断高斯分布扩展到更高维度非常直接。对于其生成模型，我们首先从 D 元正态分布中抽取一个样本 $(\nu_1, \ldots, \nu_D)$，其中 $\sim \mathcal{N}$ 服从 $(0, \Sigma)$，然后进行逐维阈值化 $\omega_i = \max\{a_i, \nu_i\}$，最终得到一个来自多元标准截断高斯分布的样本。与第一个模型一样， Σ具有单位对角元素和零位置向量，因为它的所有边缘分布都是标准截断高斯分布。 D 维联合分布的概率密度函数由此可写为:

$$\psi_D(\omega_1, \ldots, \omega_D|\Sigma, \boldsymbol{a}) = \left[\prod_{i=1}^{D} \hat{R}_i(a_i)\right] \phi_D(\nu_1, \ldots, \nu_D|\Sigma) \qquad (16)$$

其中， $\boldsymbol{a}$ 为阈值向量 a_i。由于每个算子仅作用于一个变量，因此它们彼此可交换。低维示例将在后面的式（20）中给出。

RGD 的一个关键性质是其可边缘化性。将高维（D）RGD 边缘化到任意特征子集 $S \subseteq D$ ，得到的仍然是约化子空间 S 中的 RGD。这一点可从生成模型以及多元正态分布具有类似的可边缘化性看出。约化子空间中的协方差由 $\{\Sigma_{i,j}\}_{i\in S,\ j\in S}$ 确定。我们将在后续小节中利用这一性质来构造最大似然估计。这里需要注意的是，RGD 的条件分布并非 RGD，这一点与多元正态分布不同。

4.2 使用修正高斯 Copula 的密度模型

在第二个模型中，我们提出一种与多元截断高斯分布相关联的新型 Copula。首先，我们尝试对 Copula 密度作一个粗略定义：

$$c_{RGD}(F_1(x_1),\ldots,F_D(x_D)) = \left.\frac{\psi_D(\omega_1,\ldots,\omega_D|\Sigma,\boldsymbol{a})}{\prod_{i=1}^{D}\psi_1(\omega_i|a_i)}\right|_{\omega_i=\Phi^{-1}\circ F_i(x_i)} \tag{17}$$

我们将其命名为修正高斯 Copula。变换后零膨胀数据的修正特性得到了直接建模。对于任意零值出现模式 $S \subseteq D$，我们可以确认，分母中的狄拉克 δ 函数始终会被分子中的狄拉克 δ 函数消去。因此，通过约分，我们的 Copula 密度在每个子空间上都有形式化定义。对于满足 $\boldsymbol{x}_{\bar{S}} = 0$、$\boldsymbol{x}_S > 0$ 的子空间，其 Copula 密度为

$$c_{RGD}(F_1(x_1),\ldots,F_D(x_D)) = \left.\frac{\int \mathrm{d}\boldsymbol{\nu}_{\bar{S}}\phi_D(\boldsymbol{\nu}_{\bar{S}},\boldsymbol{\omega}_S|\Sigma)}{\left[\prod_{i\in\bar{S}}\Phi(a_i)\right]\left[\prod_{j\in S}\phi_1(\omega_j)\right]}\right|_{\omega_i=\Phi^{-1}\circ F_i(x_i)} \tag{18}$$

其中，$\int \mathrm{d}\boldsymbol{\nu}_{\bar{S}}$ 是 $\prod_{i\in\bar{S}}\int_{-\infty}^{a_i}\mathrm{d}\nu_i$ 的缩写

由于原始变量 x_i 的边缘分布如式（1）所示，我们第二个模型的总体联合概率密度函数随后为

$$f_D(x_1,\ldots,x_D) = \left[\prod_{i=1}^{D}\Big(q_i\delta(x_i)+(1-q_i)\tilde{f}_i(x_i)\Big)\right]\times c_{RGD}. \tag{19}$$

在此形式下，2^D 个子空间的密度在展开乘积后，可通过 2^D 项一次性表示。

4.3 参数估计

在本小节中，我们介绍如何估计修正高斯 Copula 模型中的参数 ˜eters q_i, f_i, a_i, Σ。边缘分布中的参数很容易估计。边缘零值发生率 q_i 可根据 x_i 的经验零值率估计，而 f_i 则可再次通过核密度估计等方法，根据仅包含 $x_i > 0$的数据进行估计。修正高斯 Copula 密度中的阈值参数 a_i 可由关系式 $a_i = \Phi^{-1}(q_i)$直接得出。相关参数 Σ的估计则需要更复杂的分析。

在我们的模型中，式（19）所示形式的总概率密度函数假设变换后的分布 $\omega_i = \Phi^{-1}\circ F_i(x_i)$ 是具有协方差 Σ 和阈值 $\boldsymbol{a}$的 RGD。这里的问题是根据观测数据 ω估计多变量 RGD 的参数；下面给出这些参数的成对最大似然估计。借助 RGD 的可边缘化性，可由式（16）得到二维 RGD 的边缘概率密度函数：

$$\begin{aligned}
&\psi_2(\omega_i,\omega_j|\Sigma_{\{i,j\}},\boldsymbol{a}_{\{i,j\}})\\
&= \delta(\omega_i-a_i)\delta(\omega_j-a_j)\int_{-\infty}^{a_i}\mathrm{d}\nu_i\int_{-\infty}^{a_j}\mathrm{d}\nu_j\phi_2(\nu_i,\nu_j|\Sigma_{\{i,j\}})\\
&+ \delta(\omega_i-a_i)\int_{-\infty}^{a_i}\mathrm{d}\nu_i\phi_2(\nu_i,\omega_j|\Sigma_{\{i,j\}})\\
&+ \delta(\omega_j-a_j)\int_{-\infty}^{a_j}\mathrm{d}\nu_j\phi_2(\omega_i,\nu_j|\Sigma_{\{i,j\}})\\
&+ \phi_2(\omega_i,\omega_j|\Sigma_{\{i,j\}})
\end{aligned} \tag{20}$$

其中 $\boldsymbol{a}_{\{i,j\}} = (a_i, a_j)$ 和 $\Sigma_{\{i,j\}} = \begin{pmatrix} 1 & \Sigma_{i,j} \\ \Sigma_{i,j} & 1 \end{pmatrix}$。与第一个模型一样，$\Sigma_{i,j}$是完整模型中的一个参数，它直接出现在二维概率密度函数中，因此我们可以利用大量数据点进行估计。然而，与第一个模型不同，由于二维分布是 RGD 而非正态分布，ω_i 与 ω_j 之间的经验协方差不再是最大似然估计。此处单个数据的真实似然为式（20），其中将狄拉克 δ 函数替换为指示函数，即：

$$\begin{aligned}
&L(\omega_i,\omega_j|\Sigma_{\{i,j\}},\boldsymbol{a}_{\{i,j\}})\\
&= \begin{cases}
\int_{-\infty}^{a_i}\mathrm{d}\nu_i\int_{-\infty}^{a_j}\mathrm{d}\nu_j\phi_2(\nu_i,\nu_j|\Sigma_{\{i,j\}}) & (x_i=x_j=0)\\
\int_{-\infty}^{a_i}\mathrm{d}\nu_i\phi_2(\nu_i,\omega_j|\Sigma_{\{i,j\}}) & (x_i=0,x_j>0)\\
\int_{-\infty}^{a_j}\mathrm{d}\nu_j\phi_2(\omega_i,\nu_j|\Sigma_{\{i,j\}}) & (x_i>0,x_j=0)\\
\phi_2(\omega_i,\omega_j|\Sigma_{\{i,j\}}) & (x_i>0,x_j>0)
\end{cases}\\
&= \begin{cases}
\Phi_2(a_i,a_j|\Sigma_{\{i,j\}}) & (x_i=x_j=0)\\
\phi_1(\omega_j)\Phi(a_i|\Sigma_{i,j}\omega_j,1-\Sigma_{i,j}^2) & (x_i=0,x_j>0)\\
\phi_1(\omega_i)\Phi(a_j|\Sigma_{i,j}\omega_i,1-\Sigma_{i,j}^2) & (x_i>0,x_j=0)\\
\phi_2(\omega_i,\omega_j|\Sigma_{\{i,j\}}) & (x_i>0,x_j>0)
\end{cases}
\end{aligned} \tag{21}$$

其中 Φ_2 是二维正态分布的累积分布函数。对所有训练数据对应的式 (21) 取对数并求和，即可得到对数似然。然后，通过数值方法在开区间 (−1,1) 内最大化关于 Σ_{ij} 的总对数似然，从而得到最大似然估计。对所有成对组合重复上述成对估计，即可得到 Σ 的完整矩阵估计。

这里提出的 Σ 估计方法是对文献 [19], 中方法的高维扩展，该文献提出了一种适用于一般并列数据的通用二元 Copula 估计方法。RGD 可边缘化至二维，因此可以使用二元方法，

因此可以估计 D 维 Copula 参数。

文献 [28] 中的作者提出了在 RGD 中基于角度的成对估计方法 Σ，并给出了其误差界。然而，他们在文献中采用的非负假设（在我们的表述中，对所有 i 均对应于 $a_i < 0$ （等价地， $q_i < 1/2$））有时并不成立。他们指出，当 $a_i > 0$ 时，要将估计误差限制在固定值以下所需的数据点数会以 $\Omega(\exp(a_i^2/2))$ 的速度呈指数增长，因为大多数样本落在阈值之外，从而导致信息丢失。另一方面，在第二个模型中，我们的估计方法充分利用了所有数据，包括校正后的数据，因此即使对于 $q_i > 1/2$，我们也预计能够实现更高效的估计。我们方法的数学依据，例如误差界和所需样本量，将留待未来研究。

4.4 似然的计算

为计算 D 维空间中的完整似然值，我们首先将概率密度函数中的狄拉克 δ 函数替换为指示函数。此过程是从公式 (20) 推导公式 (21) 的推广。因此，似然值为：

$$\begin{aligned} &L(x_1, \dots x_D, x_{\bar{S}} = 0, x_S > 0) \\ &= \left[\prod_{i\in\bar{S}} q_i\right] \left[\prod_{j\in S} (1-q_j)\tilde{f}_j(x_j)\right] c_{RGD}. \end{aligned} \tag{22}$$

在实现中，式（18）定义的修正高斯 Copula 密度在 $\boldsymbol{\nu}_{\bar{S}}$ 上的积分 c_{RGD}，可使用能够数值计算以下形式的多元正态分布累积分布函数的程序库求得

$$\begin{aligned} &\int \mathrm{d}\nu_{\bar{S}} \phi_D(\boldsymbol{\nu}_{\bar{S}}, \boldsymbol{\omega}_S|\Sigma) \\ &= \phi_{|S|}(\boldsymbol{\omega}_S|\Sigma_S) \int \mathrm{d}\boldsymbol{\nu}_{\bar{S}} \phi_{D-|S|}(\boldsymbol{\nu}_{\bar{S}}|\boldsymbol{\mu}_{\bar{S}|S}, \Sigma_{\bar{S}|S}). \end{aligned} \tag{23}$$

条件均值和协方差由多元正态分布的经典结果确定，分别为 $\boldsymbol{\mu}_{\bar{S}|S} = \Sigma_{\bar{S}S}\Sigma_S^{-1}\boldsymbol{\omega}_S$和 $\Sigma_{\bar{S}|S} = \Sigma_{\bar{S}} - \Sigma_{\bar{S}S}\Sigma_S^{-1}\Sigma_{\bar{S}S}^T$。根据用于计算累积分布函数的算法，似然的计算可能需要指数级的时间，且所需时间会随着 D 的增大而增加。尽管可以使用设计良好的算法来计算多元正态分布的累积分布函数，例如 [30], 中的方法，但我们在此提供一种似然的近似计算方法，其中部分忽略相关性在零值出现之间

$$\begin{aligned} \int \mathrm{d}\boldsymbol{\nu}_{\bar{S}} \phi_D(\boldsymbol{\nu}_{\bar{S}}, \boldsymbol{\omega}_S|\Sigma) &\approx \phi_{|S|}(\boldsymbol{\omega}_S|\Sigma_S) \int \mathrm{d}\boldsymbol{\nu}_{\bar{S}} \prod_{i\in\bar{S}} \phi_1(\nu_i) \\ &= \phi_{|S|}(\boldsymbol{\omega}_S|\Sigma_S) \prod_{i\in\bar{S}} q_i. \end{aligned} \tag{24}$$

这种近似通过绕过高维数值积分，大幅降低了计算成本；计算显然可在多项式时间内完成；

$$\tilde{c}_{RGD} \approx \frac{\phi_{|S|}(\boldsymbol{\omega}_S|\Sigma_S)}{\prod_{j\in S}\phi_1(\omega_j)} = (2\pi)^{|S|/2}\phi_{|S|}(\boldsymbol{\omega}_S|\Sigma_S - I_S) \tag{25}$$

其中 I_S 是大小为 $|S|$ 的单位矩阵。尽管 $\tilde{c}_{RGD}$ 的近似形式只考虑正值变量之间的相关性，但参数 Σ 是我们使用包含一些零值元素的全部训练数据估计得到的。

我们通过阐述两个模型的异同，总结前两节中提出的方法。第一个模型针对 ZICAR 情形，其中零膨胀完全随机，与原本应取得的值无关。它也可用于完全随机缺失（MCAR）类型的情形。第二个模型假设观测到的零值背后的虚拟值低于某个阈值，因此对于现实中的零膨胀机制而言，后一个模型似乎更加自然。另一方面，第一个模型对不同变量间零值出现的相关性捕捉得更好，因为我们可以使用任意的多元二元掩码模型，例如受限玻尔兹曼机。第二个模型中的零值相关性仅通过多元正态分布的阈值化来表示：

$$\begin{aligned} &P(\boldsymbol{x}_{\bar{S}} = 0, \boldsymbol{x}_S > 0) \\ &= \int_{-0}^{+0} \mathrm{d}\boldsymbol{x}_{\bar{S}} \int_{+0}^{\infty} \mathrm{d}\boldsymbol{x}_S \; f_D(x_1, \dots, x_D) \\ &= \left[\prod_{i\in\bar{S}} q_i\right] \int_{+0}^{\infty} \mathrm{d}\boldsymbol{x}_S \; \left[\prod_{i\in S}(1-q_i)\tilde{f}_i(x_i)\right] c_{RGD}|_{\boldsymbol{x}_{\bar{S}}=0, \boldsymbol{x}_S>0} \\ &= \int_{+0}^{\infty} \mathrm{d}\boldsymbol{x}_S \left[\prod_{i\in S}(1-q_i)\tilde{f}_i(x_i)\right] \frac{\int \mathrm{d}\boldsymbol{\nu}_{\bar{S}} \phi_D(\boldsymbol{\nu}_{\bar{S}}, \boldsymbol{\omega}_S|\Sigma)}{\left[\prod_{j\in S}\phi_1(\omega_j)\right]}\Bigg|_{\omega_i = \Phi^{-1}\circ F_i(x_i)} \\ &= \left[\prod_{i\in S}\int_{a_i}^{\infty} \mathrm{d}\boldsymbol{\omega}_i\right] \left[\prod_{j\in\bar{S}}\int_{-\infty}^{a_i} \mathrm{d}\nu_{\bar{j}}\right] \phi_D(\boldsymbol{\nu}_{\bar{S}}, \boldsymbol{\omega}_S|\Sigma). \end{aligned} \tag{26}$$

每种零值出现模式的概率，都表示为对 $\mathcal{N}(0, \Sigma)$ 的密度分数在一个

由阈值 $\boldsymbol{a}$ 划分的正交象限上的积分。换言之，它等价于由 $I(\omega_i > a_i)$ 表示的 D 个二元变量，其中 $(\omega_1, \ldots, \omega_d) \sim \mathcal{N}(0, \Sigma)$。这种多元伯努利分布及其连续松弛在参考文献 [31] 中进行了讨论。由于该分布的自由度为 D^2 阶，其表达能力低于 RBM 等复杂模型。正值之间、零值之间以及正值与零值之间的相关性，都由单个修正高斯 Copula 表示，并在 ZIBT 模型中相互交织。我们的两种模型都能够处理数量呈指数级增长且维数各异的子空间。当 ZIBT 模型使用近似似然时，参数估计和似然计算都可在多项式时间内完成。

5 数值实验

我们使用合成和真实的多元零膨胀数据开展了一系列数值实验，以展示我们的密度估计模型的性能。

我们注意到，由于子空间维数不同，密度估计中广泛使用的性能指标——总对数似然——并不适用于零膨胀数据。例如，高斯混合模型可以通过将高斯分量的宽度缩小至无穷小，并将其置于低维子空间上，从而获得无限大的似然值。因此，在我们的实验中，性能是在监督式异常检测设置下进行测量的 [32]。正常测试数据集取自给定的真实分布，或从真实数据集中划分得到。将所有正常测试数据复制并加以污染以生成异常数据，随后测量异常检测任务的性能。模型输出的负对数似然与二元异常标记之间的受试者工作特征曲线下面积（AUC）值用作评估指标。在污染过程中，将正值分量 x_i 替换为从m_i 与 M_i之间的均匀分布中独立同分布抽取的样本。零值分量不进行污染。边界 m_i 和 M_i 分别由训练数据中 x_i 的第 1 和第 99 百分位点确定，其中训练数据具有 $x_i > 0$。请注意，对于更高维度 D，任务会变得更容易，因为与由 m_i 和 M_i指定的超矩形相比，数据流形更小。

5.1 合成数据中的消融研究

第一个实验使用合成数据完成。我们从图 1(a) 和 (b) 所示的两种生成过程中生成数据集。

5.1.1 数据生成过程

对于 ZICAR 数据，数据生成过程的细节如下。首先，我们随机初始化真实分布。需要初始化的参数包括边缘分布 $g_{\{i\}}$、协方差参数 Σ 和二值掩码分布。我们不随机初始化边缘分布或其累积分布函数 G_i，而是对随机严格递增函数 h_i 进行采样：$(-\infty, \infty) \to (0, 1)$，并使用 h_i 代替 $G_i^{-1} \circ \Phi$。h_i 由 Sigmoid 函数 $h_i($ $_j\sigma$ $_j\, x_i -$ c_i π_i $)$ $= \sum_{j=1}^{5} \pi_i$ 的加权和生成，其中 b_i、 (,j) 中的 $_j$ 是从参数为 (1, 1, 1, 1, 1) 的 Dirichlet 分布中抽取的权重，b_i、$_j$ 从 (0, 2) 中均匀抽取，c_i、$_j$ 从 $(-5, 5)$ 中均匀抽取。多元正态分布的协方差 Σ 从自由度为 D、矩阵参数为 I_D 的 Wishart分布中抽取。随后将 Σ 归一化为 $\Sigma_{ij} \to \Sigma_{ij}/\sqrt{|\Sigma_{ii}\Sigma_{jj}|}$，以确保对角元素均为 1。多元二元掩码分布由含有 D 个可见单元和 $2^{D/2}$ 个隐藏单元的 RBM 定义。隐藏层和交互项的权重取自 $\mathcal{N}(0, 0.12)$，可见单元的权重取自 $\mathcal{N}(1, 0.12)$。接下来，我们从固定的真实分布中抽取样本。多变量正态样本从 $\mathcal{N}(0, \Sigma)$ 中抽取，并对每个元素应用随机单调函数 h_i，使样本符合图 1(a) 中间面板所示的分布。二值掩码从 RBM 中抽取并应用于样本，从而得到图 1(a) 右侧面板所示的 ZICAR 数据。

对于 ZIBT 数据，需要初始化的参数包括˜正值数据的边缘概率密度函数 f_i、边缘零值发生率 q_i、阈值 a_i 和协方差矩阵 Σ。协方差 Σ 的生成方式与 ZICAR 情形相同。边缘零值率 q_i 从 (0, 0.5) 中均匀采样，并转换为阈值； $a_i = \Phi^{-1}(q_i)$。不连续随机非递减函数（在图 1(b)中用作 $F^{-1} \circ \Phi$ ）按以下方式采样。首先，抽取 $M = 10^4$ 个独立同分布的标准正态样本，并应用与 ZICAR 情形相同的随机严格递增函数。接着，将从均值为 $1 - q_i$的伯努利分布中抽取的 M 个独立同分布的二元样本与这些样本相乘，从而得到 M 个单变量随机零膨胀分布的样本。然后，可以根据这些样本经验性地构造一个映射，将样本映射到单变量标准 RGD，以替代 $\Phi^{-1} \circ F_i$。最后，通过反演所构造的经验映射，得到不连续随机非递减函数 h_i，即 $F_i^{-1} \circ \Phi$的一个样本。在从固定真实分布中采样数据的过程中，从 $\mathcal{N}(0, \Sigma)$ 中抽取多变量正态样本，并在各个 a_i 处进行阈值化，以生成多变量 RGD 数据，如图1(b)中间面板所示。

表 1：平均 AUC 值（ 每种合成数据在不同模型下的结果。最高值以<u>下划线</u>标示 s.

				ZICAR 模型				ZIBT 模型			
		GMM	KDE	full	w/o RBM	w/o MLE	w/o 重新缩放	full	w/ 约	w/o MLE	w/o 重新缩放
ZICAR DATA	$D=2$	0.6502	0.6871	<u>0.8091</u>	0.7983	0.7514	0.7888	0.7308	0.7353	0.7291	0.6868
	$D=5$	0.7460	0.8598	<u>0.9274</u>	0.9207	0.8483	0.9145	0.8279	0.8332	0.8292	0.7768
	$D=10$	0.8491	0.8663	<u>0.9534</u>	0.9513	0.9126	0.9489	0.8995	0.9048	0.8992	0.8554
	$D=15$	0.8719	0.7849	<u>0.9795</u>	0.9789	0.9408	0.9777	0.9310	0.9378	0.9332	0.8960
ZIBT DATA	$D=2$	0.6921	0.7032	0.7359	0.7402	0.7337	0.7375	<u>0.7717</u>	0.7587	0.7680	0.7550
	$D=5$	0.7732	0.8636	0.8165	0.8117	0.8069	0.8076	<u>0.9181</u>	0.8742	0.8901	0.9081
	$D=10$	0.8899	0.9211	0.9034	0.8915	0.9061	0.8985	<u>0.9792</u>	0.9525	0.9581	0.9786
	$D=15$	0.9381	0.9112	0.9579	0.9480	0.9574	0.9448	<u>0.9959</u>	0.9838	0.9901	0.9945

将上述构造的不连续随机单调函数h_i 应用于每个元素后，最终得到如图1(b)右侧面板所示的 ZIBT 数据。

在这两种情况下，训练数据和测试数据的大小均为 10000。测试数据的一半是从真实分布中抽取的正常数据，其余 5000 个则是通过污染重复的正常测试数据生成的异常测试数据。我们使用不同的随机种子进行了 15 次试验，计算 AUC 分数并取平均值。

5.1.2 消融

我们进行了消融研究，以检验模型各组成部分的有效性。我们使用以下后缀来区分两个模型各自经过测试的变体：

full：每个模型均采用包含所有组成部分的完整版本。正样本的边缘分布通过核密度估计进行估计，ZICAR 模型中的二值掩码则由具有 D 个可见层和 $2D$ 个隐藏层的受限玻尔兹曼机学习得到

不使用 RBM：二值掩码服从相互独立的伯努利分布，不使用 RBM。此变体仅适用于 ZICAR 模型

启用近似：似然函数使用公式 (25) 中的近似方法计算。此变体仅适用于 ZIBT 模型。禁用 MLE：使用包括零值在内的所有数据，通过经验协方差矩阵估计 Σ。禁用重新缩放：不对变量进行启发式重新缩放。

作为基线方法，我们还测量了传统密度估计模型——高斯混合模型和核密度估计——的性能。每次 15 次试验之前，均使用从初始化分布中独立生成的数据调整所有超参数。

5.1.3 结果

测得的平均 AUC 值汇总于表 1。对于每种数据生成过程（ZICAR 数据和 ZIBT 数据），我们设置了不同的维数 D。最佳分数以带下划线的形式标出。我们可以清楚地看到，完整模型在所有维数下的表现均最佳；正如预期的那样，针对 ZICAR/ZIBT 数据的最佳模型分别是完整的 ZICAR/ZIBT 模型，这证实了我们的两个完整模型确实能够捕捉真实分布。即使我们未能为数据的零膨胀机制选择合适的模型，例如错误地为 ZICAR 数据选择 ZIBT 模型，反之亦然，在大多数情况下，其表现仍优于基线方法，这表明我们的模型能够妥善处理零膨胀数据不同的维数。

对于进行消融的模型，其性能相比相应的完整模型略有下降。这有力地表明，所有组件都对模型质量有所贡献。消融模型的性能仍优于基线方法和选择不当的模型，例如，在 ZICAR 数据上，带有消融的 ZICAR 模型优于完整的 ZIBT 模型。从这些观察结果中可以得到一个重要教训：要从 ZICAR、ZIBT 或未来的其他模型中，选择一个适合数据零膨胀机制的模型，这比模型内部复杂的组件更为重要。

在 ZICAR 数据和 ZICAR 模型的消融实验中，去除 MLE 后性能下降最明显；而在 ZIBT 数据和 ZIBT 模型中，性能下降较小。为

表2：采用不同估计方法对每组数据的协方差矩阵进行估计所得的平均 L^2 估计误差。最小值以带下划线的形式标出。

		ZICAR 模型		ZIBT 模型	
		full	w/o MLE	full	w/o MLE
ZICAR DATA	$D=2$	<u>0.005</u>	0.866	0.641	0.524
	$D=5$	<u>0.052</u>	1.716	1.302	1.063
	$D=10$	<u>0.118</u>	2.658	2.042	1.665
	$D=15$	<u>0.195</u>	3.363	2.537	2.066
ZIBT DATA	$D=2$	0.190	0.112	<u>0.005</u>	0.062
	$D=5$	0.848	0.326	<u>0.037</u>	0.305
	$D=10$	1.321	0.514	<u>0.096</u>	0.421
	$D=15$	1.709	0.668	<u>0.148</u>	0.528

进一步考察 MLE 的有效性，我们汇总了每种数据和模型在 15 次试验中协方差矩阵的 L^2 估计误差，结果见表2。正如预期，适用的完整模型采用 MLE 时，估计误差最小（以下划线标示）。对于 ZICAR 数据，ZICAR 去除 MLE 的模型估计误差较大；而 ZIBT 中对应的估计误差较小，这与 ZICAR 数据上 ZICAR 去除 MLE 的模型 AUC 下降最明显这一结果一致。

5.2 真实零膨胀数据实验

第二个实验使用真实世界的多元零膨胀数据。

5.2.1 数据

具有大量零值数据的多变量零膨胀非负连续数据开放数据集相当罕见。UCI 信用数据 [33,34] 包含 30000 条记录，记录了台湾信用卡服务中个人的信用额度、性别、教育程度、婚姻状况、支付历史等信息。变量 PAY AMT{i} 和 BILL AMT{i}（ $i=1$，.....，6）表示客户之前各月的支付金额和账单金额。这 12 个变量是相关的零膨胀连续数据，零值比例约为 $6\sim24$%，且呈右偏分布。本实验使用这 12 个变量。请注意，数据中存在少量负值，已将其替换为零。随机选取的 21000 条记录用作训练数据，其余 9000 条记录用作正常测试数据。另通过污染生成 9000 条异常测试数据，方法是对重复的正常测试数据进行污染

表 3：不同模型在 UCI 信用数据上的平均 AUC 值。最大值以下划线标示。

			ZICAR 模型	ZIBT 模型	
	GMM	KDE	full	full	w/ 约
$D=2$	0.866	0.899	0.906	<u>0.917</u>	0.916
$D=12$	0.953	0.953	0.966	<u>0.987</u>	0.985

。所有超参数均在训练数据中调优。计算 15 个随机种子下 AUC 的平均值。除了使用所有零膨胀变量进行 $D=12$ 实验外，我们还进行了 $D=2$ 实验，仅使用 PAY AMT1 和 BILL AMT1，以评估低维情形下的性能。

5.2.2 结果

平均 AUC 值汇总于表 3。我们所有模型及其近似模型的表现均优于基线方法。ZIBT 完整模型在$D=2$ 和 $D=12$方面均表现最佳。这一结果表明，对于 UCI 信用数据中的零膨胀机制，ZIBT 类型的情景比 ZICAR 更为合适，这与我们的直觉一致。也就是说，零值数据并非只是信息缺失，而是表示其数值低于某个阈值。尽管通过绕过高维数值积分显著提升了计算效率，ZIBT 近似模型的性能下降幅度仍然很小。

6 结论与未来方向

多元零膨胀非负连续数据的密度估计颇具挑战性，原因在于子空间数量呈指数级增长、各子空间的维数不同，以及概率密度函数发散。本文讨论了两种零膨胀情形：ZICAR 和 ZIBT，并相应提出了两种基于 Copula 的密度估计模型，即 ZICAR 模型和 ZIBT 模型。尤其是在 ZIBT 模型中，我们提出了一种新型 Copula——修正高斯 Copula，以适应零值处的并列数据，并为 RGD 开发了最大似然估计方法。在这两种模型中，密度模型的可边缘化性使我们能够构建高效的参数估计方法，充分利用几乎所有训练数据。参

数估计和似然计算均可在多项式时间内完成。我们的模型妥善解决了多元零膨胀数据密度估计中的难题，这些难题源于子空间数量呈指数级增长、各子空间的维数不同，以及概率密度函数发散。

在对人工生成的 ZICAR 和 ZIBT 数据进行的数值实验中，我们证明了针对每种零膨胀情景恰当选择的模型都能最佳地拟合数据。我们还证实，所提出方法的所有组成部分，即受限玻尔兹曼机、最大似然估计和重新缩放技巧，都能提升模型性能。针对真实世界零膨胀数据的数值实验表明，与高斯混合模型和核密度估计等传统模型相比，我们的模型更有效。其中，ZIBT 模型表现最佳。这表明，正如我们所预期的那样，对于这些数据，ZIBT 类型的情景更为合适：在这种情景中，零值数据被视为较小的数值，而不仅仅是信息缺失。

在多元零膨胀数据的密度估计中，需要捕捉正值之间的相关性、零值出现之间的相关性，以及正值与零值之间的相关性。我们的 ZICAR 模型无法捕捉第三种相关性；而在 ZIBT 模型中，所有类型的相关性都同时通过单个修正高斯 Copula 表示，因此彼此纠缠。我们可能需要更灵活、表达能力更强的多元零膨胀数据建模工具。

致谢

作者谨向 S. Matsumoto 致以谢意，感谢其对稿件早期草稿提出的宝贵意见，并感谢 T. Yoshiba 和 M. Kazato 富有成效的讨论。

参考文献

[1] Christopher M Bishop 和 NasserM Nasrabadi。《模式识别与机器学习》，第 4 卷。施普林格，2006 年。

[2] Emanuel Parzen。《概率密度函数与众数的估计》。《数学统计年刊》，33(3)：1065–1076，1962。[3] Juliana Schulz、Christian Genest 和 Mhamed Mesfioui。《基于共单调冲击的多元泊松模型》。《国际统计评论》，89(2)：323–348，2021年8月。

[4]Hossein Pishro-Nik。《概率、统计学与随机过程导论》。2016年。

[5] D. Lambert。零膨胀泊松回归及其在制造缺陷中的应用。《计量学》，34(1)：1–14，1992 年 2 月。

[6]Martin Ridout、John Hinde 和 Clarice G. B. Dem´etrio。检验零膨胀泊松回归模型与零膨胀负二项分布替代模型的得分检验。《生物计量学》，57(1)：219–223，2001 年 3 月。

[7] Lei Liu、Ya-Chen Tina Shih、Robert L. Strawderman、Daowen Zhang、Bankole A. Johnson 和 Haitao Chai。《零膨胀非负连续数据的统计分析：综述》。《统计科学》，34(2)，2019年5月。

[8] David W Scott。《多元密度估计：理论、实践与可视化》。John Wiley & Sons，2015年。

[9] RogerB Nelsen。《Copula 导论》。施普林格科学与商业媒体，2007年。

[10] Harry Joe。具有给定边缘分布和 m(m-1)/2 个二元依赖参数的 m元分布族。《讲义—专著系列》，第120–141页，1996年。

[11]Tim Bedford 和 Roger M. Cooke。藤结构——一种用于相关随机变量的新型图形模型。《统计年鉴》，30(4)，2002年8月。

[12] Kjersti Aas、Claudia Czado、Arnoldo Frigessi 和 Henrik Bakken。《多重依赖的配对 Copula 构造》。《保险数学与经济学》，44(2):182–198，2009年4月。

[13] Eric Bouy´e、Valdo Durrleman、Ashkan Nikeghbali、Ga¨el Riboulet 和 Thierry Roncalli。《金融Copula——阅读指南及若干应用》。可通过 SSRN 1032533 获取，2000年。

[14] Andrew J. Patton。《经济时间序列 Copula模型综述》。《多元分析杂志》，110:4–18，2012年9月。

[15] Guanqi Fang、Rong Pan 和 Yili Hong。《基于 Copula 的退化系统相关失效可靠性分析》。《可靠性工程与系统安全》，193:106618，2020年1月。

[16]Pierre Georges、Arnaud-Guilhem Lamy、Emeric Nicolas、Guillaume Quibel 和 Thierry Roncalli。

多变量生存建模：一种基于 Copula 的统一方法。可通过 SSRN 1032559 获取，2001 年。

[17]Tomasz Burzykowski、Geert Molenberghs、Marc Buyse、Helena Geys 和 Didier Renard。多项以失效时间为终点的随机临床试验中替代终点的验证。《皇家统计学会会刊》C辑（应用统计学），50(4):405–422，2001年1月。

[18]Ivan Kojadinovic。针对存在并列值的情形调整的若干 Copula 推断程序。《计算统计学与数据分析》，112:24–41，2017年8月。

[19] Yan Li、Yang Li、Yichen Qin 和 Jun Yan。含并列值数据的 Copula 建模。《统计学及其界面》，13(1):103–117，2020年。

[20] David C Howell。缺失数据的处理。《Sage社会科学方法论手册》，第208–224页，2007年。

[21] Hyun Kang。缺失数据的预防与处理。《韩国麻醉学杂志》，64(5):402，2013年。

[22] Roderick JA Little 和 Donald B Rubin。《缺失数据的统计分析》，第 793 卷。John Wiley & Sons，2019 年。

[23]Yoav Freund 和 David Haussler。使用双层网络对二进制向量上的分布进行无监督学习。载于 J. Moody、S. Hanson 和 R.P. Lippmann 编，《神经信息处理系统进展》第 4 卷。Morgan-Kaufmann，1991 年。

[24] Miguel ´A. Carreira-Perpi˜n´an 和 Geoffrey Hinton。关于对比散度学习。载于 Robert G. Cowell 和 ZoubinGhahramani 编，《第十届国际人工智能与统计学研讨会论文集》，《机器学习研究论文集》第 R5 卷，第 33–40 页。机器学习研究论文集（PMLR），2005 年 1 月 6–8 日。2021 年 3 月 30 日由机器学习研究论文集（PMLR）重新发布。

[25] James Tobin。受限因变量关系的估计。《计量经济学》：计量经济学学会期刊，第 24–36 页，1958 年。

[26] David G. Kleinbaum 和 Mitchel Klein。《生物学与健康统计学》。Springer纽约，2012 年。

[27] Nicholas Socci、Daniel Lee 和 H. Sebastian Seung。《截断高斯分布》。载于 M. Jordan, M. Kearns 和 S. Solla 编，《神经信息处理系统进展》第 10 卷。MIT 出版社，1997 年。

[28]Shanshan Wu、Alexandros G Dimakis 和 Sujay Sanghavi。学习由单层 ReLU 网络生成的分布。载于 H. Wallach、H. Larochelle、A. Beygelzimer、F. d'Alch´e-Buc、E. Fox 和 R. Garnett 编，《神经信息处理系统进展》，第 32 卷。Curran Associates, Inc.，2019 年。

[29] Markus Harva 和 Ata Kab´an。整流因子分析的变分学习。《信号处理》，87(3)：509–527，2007 年 3 月。

[30] Alexandra Gessner、Oindrila Kanjilal 和 Philipp Hennig。线性域约束下的高斯分布积分。载于《国际人工智能与统计学会议》，第2764–2774页。机器学习研究论文集，2020年。

[31] Xi Wang 和 Junming Yin。松弛的多元伯努利分布及其在深度生成模型中的应用。载于 Jonas Peters 和 David Son- tag 编，UAI 第36届人工智能不确定性会议论文集，《机器学习研究论文集》第124卷，第500–509页。机器学习研究论文集，2020年8月3–6日。

[32]Andrew Emmott、Shubhomoy Das、Thomas Dietterich、Alan Fern 和 Weng-Keen Wong。异常检测问题的元分析。arXiv 预印本 arXiv:1503.01158，2015年。

[33] Dheeru Dua 和 Casey Graff。《UCI 机器学习资料库》，2017 年。

[34] I-Cheng Yeh 和 Che-hui Lien。《信用卡客户违约概率预测准确率的数据挖掘技术比较》。《专家系统与应用》，36(2):2473–2480，2009 年。