arXiv:2601.05345v1 [stat.ME] 8 Jan 2026

使用新型圆回归混合模型进行基于模型的聚类

SphiweBonakeleSkhosana* 和 Najmeh Nakhaei Rad

比勒陀利亚大学统计学系，南非比勒陀利亚，0028

*通讯作者： spiwe.skhosana@up.ac.za

2026年1月12日

摘要

响应变量为圆形变量的回归模型，在生物学、地质学和气象学等领域十分常见。典型模型假设响应的条件分布服从冯·米塞斯分布。然而，当响应变量呈多峰性时，这一假设并不适用。因此，本文针对圆形响应变量以及一组圆形和／或线性协变量，提出一种有限回归混合模型。当底层总体呈多峰性时，混合模型非常有用。尽管在圆形数据的回归建模中多峰性十分普遍，文献中却鲜有研究采用回归混合模型。本文旨在弥补这一知识空白。为估计所提出的模型，我们通过期望最大化算法开发了一种极大似然估计程序。我们通过一项广泛的模拟研究，展示所提出模型及其估计程序的实际应用价值和性能。此外，我们还证明，该模型可用作基于模型的聚类工具。最后，我们将该模型应用于来自南非某风电场的真实数据集。

关键词：*von-Mises*混合模型；圆形数据；基于模型的聚类；期望-最大化算法；圆周回归

1 引言

具有“方向性”属性（角度）的数据称为圆形数据，这类数据被认为较为特殊，由于其周期性，无法使用笛卡尔坐标系进行建模。因此，有必要发展超越传统分析框架的统计技术，以便在众多科学领域中有效地对此类数据进行建模、解释和利用。当响应变量（记为 θ）在圆上测量，而协变量$\mathbf{x} \in \mathbf{R}^p$ 在实数轴上测量时，就会产生圆形-线性回归模型。这类情况常见于生物学、地质学和气象学等诸多科学领域。例如，为了开展保护研究，可以研究动物移动方向与其移动距离之间的依赖关系（Fisher and Lee [1992]），或研究鱼类产卵时间与低潮振幅之间的依赖关系（例如Lund [1999]）。关于地质学和气象学中的应用，请参见Fisher and Lee [1992] 以及Jammalamadaka 和 Sengupta [2001]中提到的示例。

圆形-线性回归模型最初由 Gould [1969] 提出，其假设误差服从冯·米塞斯分布。Johnson和Wehrly [1978] 指出，Gould 模型存在一个严重问题：回归函数会像旋转的“理发店灯柱”一样，在圆柱体上无限次螺旋缠绕。这意味着该模型的似然函数存在无穷多个同等重要的极大值，导致模型无法识别，因此在实际应用中没有用处。对于单个协变量，作者（Johnson和Wehrly [1978]）改进了 Gould 模型，提出一种回归函数在圆柱体上只完成一次螺旋缠绕的模型，从而得到唯一的模型和唯一的解。 Fisher and Lee [1992] 将 Johnson-Wehrly 模型推广到包含多个协变量的情形，并进一步扩展了模型，使圆形响应的离散程度可以依赖于协变量。关于圆形-线性回归模型，以及一般而言圆形或方向数据的回归模型，更多详情请参见（Fisher and Lee [1992]；第 6.4 节）和（Mardia 和 Jupp [2000]；第 11.3 节）。

越来越多的圆周数据集呈现出多峰性等非平凡特征（Ley和Verdebout [2017]）。当圆形响应 θ 呈现多峰性时，圆形-线性回归模型并不是一个恰当的模型，而这种情况在实践中很常见。例如，在上述动物定向示例中，动物可能必须在两个运动方向之间进行选择，这两个方向都较为偏好Batschelet [1981]。在

一个24小时周期中，某种鱼类的产卵高峰时刻可能出现在傍晚的早些时候（例如下午6点左右），也可能出现在午夜（例如凌晨12点左右）（Child等 [1991]）。呈现此类模式的样本数据可以被认为是由两个（或更多）单峰分布生成的。与上述示例不同，在实践中的大多数情况下，多个方向（或时间峰值）之间的角度（或时间）可能是任意的，没有标准方法将观测数据分成单峰数据样本（Batschelet[1981]）。在圆形或更一般的方向数据分析中，处理多峰性的一种客观方法是使用混合模型（Fisher等 [1993] 和 McLachlan和Peel [2000]）。

混合模型适用于对来自某个总体的样本数据进行建模；该总体由例如 K 个未知子总体构成，此时这类模型称为有限混合模型（Fr¨uhwirth-Schnatter [2006]）。每个子总体都可以具有其独特的概率分布。有限混合模型早期的应用包括 Newcomb（Newcomb[1886]）和 Pearson（Pearson [1894]）分别在天文学和动物学中的研究。后来，有限混合模型的应用扩展到金融学（用于对金融回报建模，Kon [1984]）、经济学（用于对收入分布建模，Schneider and Scharfenaker[2020]）、生态学（用于对物种分布建模，Joseph et al. [2009]）等领域。有限混合模型在圆形数据领域的应用包括 Stephens[1969]，他除其他研究外，还将一个含有 $K = 2$ 个成分的冯·米塞斯分布混合模型拟合到海龟运动数据。此后，冯·米塞斯分布混合模型被用于对婴儿猝死综合征的季节性建模（Mooney et al. [2003]）、对风向建模（Carta et al. [2008] 和 Masseran et al. [2013]），以及机器人学中的仅方位跟踪（Markovi´c and Petrovi´c [2012]）。最近，Jammalamadaka and Vaidyanathan[2024]提出了一种冯·米塞斯分布混合模型，用于圆形数据聚类。为处理数据中的离群值，Bentley [2006]提出了冯·米塞斯分布与均匀分布的混合模型。为处理数据中的偏度，Rad et al. [2022] 提出了正弦偏斜冯·米塞斯分布混合模型，用于对风向建模。

一般有限混合模型中一个极具实用价值的类别是有限回归混合模型。由于具有灵活性，有限回归混合模型允许每个子总体在响应变量与一组协变量之间拥有独特的关系。有限回归混合模型起源于经济学中的切换回归模型（Quandt [1972], Goldfeld andQuandt [1973] 和Quandt and Ramsey [1978]）。其

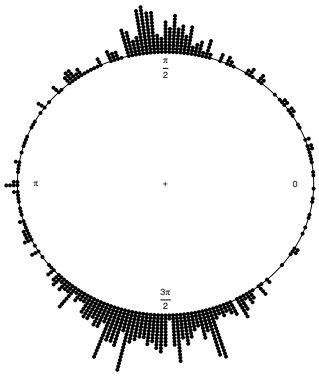
在经济学中的成功促使其被广泛应用于市场营销（DeSarboand Cron [1988] 作为潜在类别回归模型）、机器学习（Jacobs etal. [1991] 作为专家混合模型）和生物学等众多应用领域。在这些应用领域中，有限回归混合模型支撑着各种统计建模技术，例如稳健回归建模（Box and Tiao [1968]）、基于模型的聚类（Ingrassiaet al. [2014]）、分布回归（Kneib et al. [2023]）、混合效应建模（Fr¨uhwirth-Schnatter [2006]；第8.5节）以及删失数据建模（Zeller et al.[2019]）。关于有限回归混合模型及一般有限混合模型的综合性综述，可分别参见（Fr¨uhwirth-Schnatter [2006] 的第8章以及Yaoand Xiang [2024]的第3章）和（McLachlan and Peel [2000]以及Fruhwirth-Schnatter et al. [2019]）。

遗憾的是，文献中尚未关注将有限回归混合模型用于圆形数据。因此，这一知识空白亟待填补，填补后将丰富有限混合建模和圆形数据分析领域的文献。本文提出一种有限回归混合模型，以下称为有限圆形回归混合模型。本文的研究动机来自一项探究影响风向因素的研究。数据包括南非东开普省一座风电场在2019年1月1日至2019年1月31日期间每10分钟记录一次的4464个观测值。在此期间，测量了以下变量：平均风向（弧度）、平均风速（米/秒）、平均气温（单位为°C），以及测量日期和时间。为了在本研究中将24小时制的时间视为圆形变量，我们决定使用小时数据。为此，我们对数据进行平均，得到小时数据。使用圆形平均方向计算每小时的平均风向，并使用常规算术平均值计算每小时的平均风速和气温（Jammalamadaka 和 Lund [2006]）。最终得到的小时数据集包含744个观测值。为评估一天中的时间对平均风向的影响，我们将一天中的小时乘以（$2\pi/24$），将其转换为圆形变量。因此，数据包含两个线性协变量（风速和气温）以及一个圆形协变量（一天中的小时）。图1a展示了平均风向（弧度）的圆形图。观测到的平均风向明显呈双峰分布，其中一个峰的平均（风）方向为

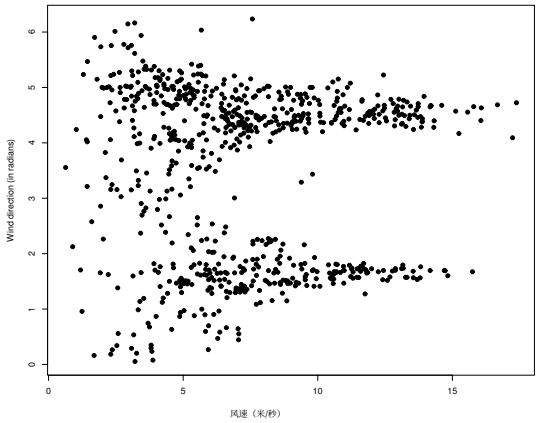
方向为90度（或 $\pi/2$弧度），另一个平均（风向）约为270度（或 $3\pi/2$弧度）。这种双峰性还明显体现在

图1b–1d所示的风向分别与风速、气温和一天中的小时的散点图中。因此，正如上文所述（更多详情见第2.2节），圆回归模型并不适用于这些数据，因为它假设圆形响应是单峰的。

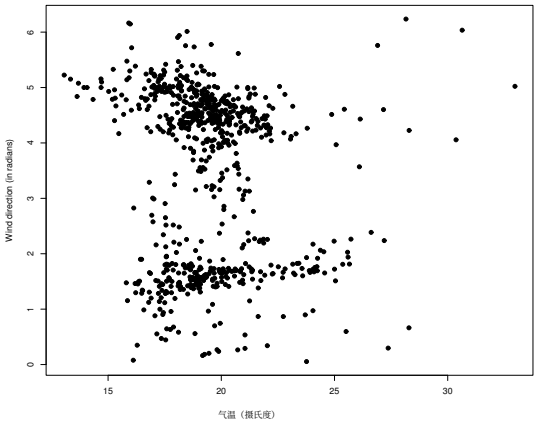
本文其余部分的结构如下。第 2 节定义 von-Mises 分布、圆形-线性回归模型及其估计程序，随后给出圆形-线性回归的有限混合及其估计程序的定义。本节最后讨论一些建模和计算方面的细节。第 3 节开展蒙特卡洛模拟，以展示所提出模型的应用以及所提出估计程序的性能。第 4 节将所提出的模型应用于风数据，以展示其在实际应用中的潜力。第 5 节总结全文，并提出未来研究方向。



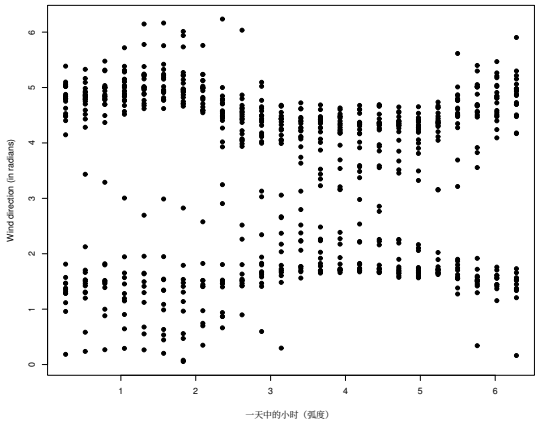
(a)



(b)



(c)



(d)

图1: （a）744个每小时平均风向的圆形图，测量单位为弧度。（b）平均风向与平均风速的散点图，风速单位为米/秒。（c）平均风向与平均气温的散点图，气温单位为摄氏度。（d）平均风向与一天中的小时的散点图，时间单位为弧度。数据记录于南非东开普省的一座风电场。

2 方法论

2.1 冯·米塞斯分布

与实数轴上连续变量所服从的高斯（正态）分布类似，冯·米塞斯分布是圆周上变量最广泛适用的概率分布（Mardia 和 Jupp [2000]）。对于单变量圆形随机变量 θ，冯·米塞斯分布的密度函数记作$\mathcal{VM}(\theta|\mu,\kappa)$，其形式为

$$\mathcal{VM}(\theta|\mu,\kappa)=\frac{1}{2\pi I_0(\kappa)}\exp(\kappa\cos(\theta-\mu)),\qquad 0<\theta\leq 2\pi, \tag{1}$$

其中 $\mu\in(0,2\pi]$ 表示平均方向，$\kappa\geq 0$ 表示集中参数。式（1）中的函数 $I_0(\kappa)$ 是零阶第一类修正贝塞尔函数（更多详情请参见 Mardia 和 Jupp [2000]及其中的参考文献）。

冯·米塞斯分布具有一些有趣的性质。当 $\kappa>0$时，密度函数（1）是单峰的，且关于 μ 对称。当 $\kappa=0$时，密度函数（1）退化为圆周上的均匀密度函数。最后，当 $\kappa\to\infty$时，密度函数（1）近似为高斯密度函数（Bishop [2006]）。因此，冯·米塞斯分布通常被称为圆周上连续变量的高斯分布，或简称圆形高斯分布（Jammalamadaka 和 Sengupta [2001]和 Bishop [2006]）。有关冯·米塞斯分布的更多详情，请参见Mardia 和 Jupp [2000] 的第 3 章。

2.2 圆周回归模型

设 $\theta\in(0,\ 2\pi]$，$\mathbf{x}_1=(x_1,\ x_2,\ \ldots,\ x_q)^\intercal\in(0,\ 2\pi]^q$ 以及$\mathbf{x}_2=(x_1,\ x_2,\ \ldots,\ x_p)^\intercal\in\mathbb{R}^p$ 分别表示圆形响应变量、圆形协变量向量和线性协变量向量。圆回归模型有三种可能的形式：(1) 仅包含线性协变量的模型，下文称为圆形-线性回归模型；(2) 仅包含圆形协变量的模型，下文称为圆-圆回归模型；(3) 同时包含圆形和线性协变量的模型，下文称为圆形-圆形-线性回归模型。下面简要讨论每种模型。

2.2.1 圆形-线性回归模型

仅给定$\mathbf{x}_2$, 时，假设 θ 的条件分布服从冯·米塞斯分布，$\mathcal{VM}(\mu(\mathbf{x}_2)\ \kappa)$，且条件平均方向 $\mu(\mathbf{x}_2)$由下式给出

$$\mu(\mathbf{x}_2) = \mu + g(\mathbf{x}_2, \boldsymbol{\beta}_2), \tag{2}$$

其中 $\boldsymbol{\beta}_2 = (\beta_1,\ \beta_2,\ \ldots,\ \beta_p)^\intercal$ 是回归系数向量，而 $g(\cdot)$ 是一个将实数轴映射到圆上的连接函数（$-\pi,\ \pi$）。

Gould [1969] 使用了恒等连接函数，由此得到的模型中，似然函数存在无穷多个具有相同似然值的极大值（Mardia 和 Jupp[2000]）。实际上，这意味着 $\boldsymbol{\beta}_2$ 没有唯一解，因而不可识别。因此，该模型在实际应用中并不实用。

通过使用满足 $g\ \ g(0) = 0$ (Fisher and Lee [1992]) 条件的单调连接函数，可以避免 Gould 模型所遇到的问题。一种可选的连接函数是反正切连接函数

$$g(\mathbf{x}_2, \boldsymbol{\beta}_2) = 2\tan^{-1}(\mathbf{x}_2^\intercal \boldsymbol{\beta}_2). \tag{3}$$

关于 g的其他可选形式，参见 FisherandLee [1992] 和Mardia 和 Jupp [2000]。

2.2.2 圆形-圆形回归模型

给定仅有 $\mathbf{x}_1$, 时，假设 θ 的条件分布服从冯·米塞斯分布，$\mathcal{VM}(\mu(\mathbf{x}_1)\ \kappa)$，且条件平均方向 $\mu(\mathbf{x}_1)$由下式给出

$$\begin{aligned} \mu(\mathbf{x}_1) &= \mu + \sum_{j=1}^{q} (\beta_{1j}\cos(x_j) + \beta_{2j}\sin(x_j)) \\ &= \mu + \mathbf{x}_1^{*\intercal}\boldsymbol{\beta}_1^*, \end{aligned} \tag{4}$$

其中 $\mathbf{x}_1^* = (\cos(x_1),\ \sin(x_1),\ \cos(x_2),\ \sin(x_2) \ldots\ldots,\ \cos(x_q),\ \sin(x_q))$，且 $\boldsymbol{\beta}_1^* = (\beta_{11}, \beta_{21}, \beta_{12}, \beta_{22}, \ldots, \beta_{1q}, \beta_{2q})$。

请注意，正如 Kim 和 SenGupta [2018], 所指出的，回归函数（4）的值域可能不在单位圆上。因此，我们提出使用反正切链接函数对（4）进行建模

$$\mu(\mathbf{x}_1) = \mu + 2\tan^{-1}(\mathbf{x}_1^{*\intercal}\boldsymbol{\beta}_1^*). \tag{5}$$

圆-圆回归模型的其他形式由 Sarma 和 Jammala- madaka [1993] 以及Downs 和 Mardia [2002], 提出，参见 Kim 和 SenGupta [2018] 的近期综述。

2.2.3 圆形-圆形-线性回归模型

给定 $\mathbf{x}_1$ 和 $\mathbf{x}_2$, ，假设 θ 的条件分布服从冯·米塞斯分布， $\mathcal{VM}(\mu(\mathbf{x}_1, \mathbf{x}_2)\ \kappa)$。条件平均方向 $\mu(\mathbf{x}_1, \mathbf{x}_2)$由公式 (2) 和 (5) 组合得到：

$$\mu(\mathbf{x}_1, \mathbf{x}_2) = \mu + g(\mathbf{x}_1^{*\intercal}\boldsymbol{\beta}_1) + g(\mathbf{x}_2^{\intercal}\boldsymbol{\beta}_2), \tag{6}$$

其中 $g(x) = 2\tan^{-1}(x)$。

这意味着，圆形-圆形-线性回归模型将前两个模型作为特例。因此，本文其余部分将聚焦于圆形-圆形-线性回归模型。为简化术语，下文将该模型称为圆回归模型。此外，为简化符号，令 $\mathbf{x} = (\mathbf{x}_1,\mathbf{x}_2)^{\intercal}$,$\mathbf{x}^* = (\mathbf{x}_1^*,\mathbf{x}_2)^{\intercal}$ 以及 $\boldsymbol{B} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$。

2.3 极大似然估计

本节将介绍式（6）中圆回归模型的估计程序。式（2）和式（5）中另外两个模型的估计程序可作为特例得到。考虑来自条件分布 $\mathcal{VM}(\mu(\mathbf{x})\ \kappa))$ 的一个随机样本 $\{(\mathbf{x}_i, \theta_i) : i = 1, 2, \ldots, n\}$，其中 $\mu(\mathbf{x})$ 由式（6）给出。对数似然函数为

$$\ell(\boldsymbol{\vartheta}|\boldsymbol{\theta}, \mathbf{x}^*) = -n\log(2\pi) - n\log(I_0(\kappa)) + \kappa\sum_{i=1}^{n}\cos(\theta_i - \mu - g(\mathbf{x}_i^{*\intercal}\boldsymbol{B})), \tag{7}$$

其中，$\boldsymbol{\vartheta} = (\mu、\ \kappa、\ \boldsymbol{B})$ 是所有模型参数组成的向量。
μ 的极大似然估计（MLE）通过关于 μ 最大化式（7）得到：

$$\frac{\partial\ell(\boldsymbol{\vartheta}|\boldsymbol{\theta}, \mathbf{x}^*)}{\partial\mu} = -\kappa\sum_{i=1}^{n}\sin(\theta_i - \mu - g(\mathbf{x}_i^{*\intercal}\boldsymbol{B})). \tag{8}$$

令式（8）等于零，得到

$$\sum_{i=1}^{n}\sin(\theta_i - \mu - g(\mathbf{x}_i^{*\intercal}\boldsymbol{B})) = 0. \tag{9}$$

利用三角恒等式

$$\sin(x - y) = \sin(x)\cos(y) - \sin(y)\cos(x), \tag{10}$$

我们求解 (9)，得到 μ

$$\hat{\mu} = \tan^{-1}\left[\frac{\sum_{i=1}^{n}\sin(\theta_i - g(\mathbf{x}_i^{*\intercal}\boldsymbol{B}))}{\sum_{i=1}^{n}\cos(\theta_i - g(\mathbf{x}_i^{*\intercal}\boldsymbol{B}))}\right]. \tag{11}$$

κ 的最大似然估计通过关于 κ最大化（7）得到，如下所示:

$$\frac{\partial \ell(\boldsymbol{\vartheta}|\boldsymbol{\theta}, \mathbf{x}^*)}{\partial \kappa} = -nA(\kappa) + \sum_{i=1}^{n}\cos(\theta_i - \mu - g(\mathbf{x}_i^{*\intercal}\boldsymbol{B})), \tag{12}$$

其中， $A(\kappa) = I_1(\kappa)/I_0(\kappa)$，且 $I_1(\kappa) = I_0^{'}(\kappa)$。

令（12）等于零并求解 κ ，得到

$$\hat{\kappa} = A^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\cos(\theta_i - \hat{\mu} - g(\mathbf{x}_i^{*\intercal}\boldsymbol{B}))\right), \tag{13}$$

其中， $A^{-1}(x)$是函数 $A(x)$的逆函数。

回归系数 $\boldsymbol{B}$ 的最大似然估计没有闭式表达式。因此，我们使用上升算法（例如牛顿-拉夫森（NR）算法）对其进行近似。 $\boldsymbol{B}$ 的NR更新规则为

$$\boldsymbol{B}^{\text{new}} = \boldsymbol{B}^{\text{old}} - \mathbf{H}^{-1}(\boldsymbol{B}^{\text{old}})\boldsymbol{g}(\boldsymbol{B}^{\text{old}}), \tag{14}$$

其中

$$\boldsymbol{g}(\boldsymbol{B}) \equiv \frac{\partial \ell(\boldsymbol{\vartheta}|\boldsymbol{\theta}, \mathbf{x}^*)}{\partial \boldsymbol{B}} = \mathbf{X}^{*\intercal}\dot{\boldsymbol{G}}\mathbf{u}, \tag{15}$$

$$\mathbf{H}(\boldsymbol{B}) \equiv \frac{\partial^2 \ell(\boldsymbol{\vartheta}|\boldsymbol{\theta}, \mathbf{x}^*)}{\partial^2 \boldsymbol{B}} = -\mathbf{X}^{*\intercal}\dot{\boldsymbol{G}}^2\boldsymbol{V}\mathbf{X}^* + \mathbf{X}^{*\intercal}\ddot{\boldsymbol{G}}\mathbf{U}\mathbf{X}^*, \tag{16}$$

with

$$\begin{aligned}
\dot{\boldsymbol{G}} &= \text{diag}\{g^{'}(\mathbf{x}_1^{*\intercal}\boldsymbol{B}), g^{'}(\mathbf{x}_2^{*\intercal}\boldsymbol{B}), \ldots, g^{'}(\mathbf{x}_n^{*\intercal}\boldsymbol{B})\},\\
\ddot{\boldsymbol{G}} &= \text{diag}\{g^{''}(\mathbf{x}_1^{*\intercal}\boldsymbol{B}), g^{''}(\mathbf{x}_2^{*\intercal}\boldsymbol{B}), \ldots, g^{''}(\mathbf{x}_n^{*\intercal}\boldsymbol{B})\},\\
\mathbf{X}^* &= (\mathbf{x}_1^{*\intercal}, \mathbf{x}_2^{*\intercal}, \ldots, \mathbf{x}_n^{*\intercal})^{\intercal},\\
\mathbf{u} &= (u_1, u_2, \ldots, u_n),\\
\mathbf{U} &= \text{diag}\{u_1, u_2, \ldots, u_n\},\\
\mathbf{V} &= \text{diag}\{v_1, v_2, \ldots, v_n\},\\
u_i &= \sin(\theta_i - \hat{\mu} - g(\mathbf{x}_i^{*\intercal}\boldsymbol{B})),\\
v_i &= \cos(\theta_i - \hat{\mu} - g(\mathbf{x}_i^{*\intercal}\boldsymbol{B})) \text{ for } i = 1, 2, \ldots, n.
\end{aligned}$$

请注意， $g'(x)$ 和 $g''(x)$ 分别是函数 $g(x)$ 的一阶导数和二阶导数。

$\boldsymbol{B}$的最大似然估计，记为$\hat{}\boldsymbol{B}$，是 NR 算法收敛时 $\boldsymbol{B}^{\text{old}}$ 的更新值。请注意，为了获得$\hat{}\mu$和$\hat{}\kappa$，我们需要知道 $\boldsymbol{B}$。按照 Fisher and Lee [1992], 中的建议，我们使用迭代方案求得$\hat{}\mu$、$\hat{}\kappa$ 和 $\boldsymbol{B}$ 。首先取 $\boldsymbol{B}$的一个初始值，记为 $\boldsymbol{B}^{(0)}$，并用它求得$\hat{}\mu$，进而得到$\hat{}\kappa$。然后利用$\hat{}\mu$，通过应用式（14）更新 $\boldsymbol{B}^{(0)}$，得到$\hat{}\boldsymbol{B}$。我们使用$\hat{}\boldsymbol{B}$ 更新$\hat{}\mu$ 和$\hat{}\kappa$，并重复这一循环直至收敛。上述估计程序总结于算法1中。

算法1 拟合圆形-线性回归模型

1: 令 $\boldsymbol{B}^{(0)}$ 为 $\boldsymbol{B}$适当选取的初始值。2: 给定$\hat{}\boldsymbol{B}^{(r-1)}$，在第 r^{th} 次迭代中，使用式（11）和式（13）估计 μ 和 κ。3: 给定步骤2得到的$\hat{}\mu^{(r)}$，使用式（14）更新 $\boldsymbol{B}^{(r-1)}$。4: 重复步骤2和步骤3，直至收敛。

2.4 圆形回归混合模型

假设随机样本 $\{\theta_1, \theta_2, \ldots, \theta_n\}$ 取自一个总体，该总体由 K 个子总体或成分组成，分别为 $1, 2, \ldots, K$；每个子总体都包含未知比例 $\pi_1, \pi_2, \ldots, \pi_K$ 的数据点，这些数据点由一个 $\mathcal{VM}(\mu_k, \kappa_k)$ 子总体生成，其中 $k = 1, 2, \ldots, K$。因此， θ 被认为是由冯·米塞斯分布的混合生成的

$$\mathcal{MVM}(\theta) = \sum_{k=1}^{K} \pi_k \mathcal{VM}(\mu_k, \kappa_k), \tag{17}$$

其中 $\pi_k > 0$（ $k = 1, 2, \ldots, K$），且满足 $\sum_{k=1}^{K} \pi_k = 1$，为混合比例。参数 μ_k 和 κ_k 分别是第 k^{th} 个成分的平均方向和集中度。

假设除了 θ_i之外，我们还观测到一个协变量向量$\mathbf{x}_i = (\mathbf{x}_{1i}, \mathbf{x}_{2i}) \in (0, 2\pi]^q \times \mathbb{R}^p$，其中 $i = 1, 2, \ldots, n$；这里，$\mathbf{x}_1$ 表示一个由 q 圆形协变量构成的向量，而$\mathbf{x}_2$ 表示一个由 p 线性协变量构成的向量。为了对 θ 与$\mathbf{x}$之间的依赖关系建模，我们提出一种圆回归混合模型，其中圆形响应变量的条件分布

θ在给定$\mathbf{X}=\mathbf{x}$的条件下，由冯·米塞斯分布混合给出

$$\mathcal{MVM}(\theta|\mathbf{X}=\mathbf{x})=\sum_{k=1}^{K}\pi_k\mathcal{VM}(\mu_k+g(\mathbf{x}^{*\intercal}\boldsymbol{B}_k)),\kappa_k), \tag{18}$$

其中 $\mathbf{x}^*=(\mathbf{x}_1^*,\ \mathbf{x}_2)^\intercal$，其中 $\mathbf{x}_1^*=(\cos(x_1),\ \sin(x_1),\ \cos(x_2),\ \sin(x_2)\ldots,\ \cos(x_q),\ \sin(x_q))$，而$\boldsymbol{B}_k$ 是 k^{th} 成分的回归系数向量。

本文假设 K 为固定值。因此，模型（18）是圆周回归有限混合模型。然而，在大多数实际应用中， K 未知，必须根据数据进行推断。在第2.6.2节中，我们提出一种数据驱动的方法来选择 K。

2.5 通过期望最大化算法进行最大似然估计

考虑来自模型（18）的随机样本 $\{(\mathbf{x}_i,\ \theta_i):\ i=1,\ 2,\ \ldots,\ n\}$，并定义$\mathbf{x}_i^*$，对于$i=1,\ 2,\ \ldots,\ n$。对数似然函数为

$$\ell(\boldsymbol{\vartheta}|\boldsymbol{\theta},\mathbf{x}^*)=\sum_{i=1}^{n}\log\sum_{k=1}^{K}\pi_k\mathcal{VM}(\mu_k+g(\mathbf{x}_i^{*\intercal}\boldsymbol{B}_k),\kappa_k), \tag{19}$$

其中 $\boldsymbol{\vartheta}=((\pi_1,\ \mu_1,\ \kappa_1,\ \boldsymbol{B}_1),\ (\pi_2,\ \mu_2,\ \kappa_2,\ \boldsymbol{B}_2)\ldots,\ (\pi_K,\ \mu_K,\ \kappa_K,\ \boldsymbol{B}_K))$是所有模型参数构成的向量。

众所周知，无法通过解析方式最大化形如 (19) 的对数似然函数，以获得 $\boldsymbol{\vartheta}$的最大似然估计。因此，我们使用期望最大化（EM）算法（Dempster等人 [1977]）对 (19) 进行数值最大化。为此，对于每个（$\mathbf{x}_i^*,\theta_i$）， $i=1,\ 2,\ldots,n$，我们定义潜变量$\mathbf{z}_i=(z_{i1},z_{i2},\ldots,z_{iK})$，使得

$$z_{ik}=\begin{cases}0 & \text{if } (\mathbf{x}_i^*,\theta_i) \text{ is not generated from the } k^{th} \text{ component,}\\ 1 & \text{if } (\mathbf{x}_i^*,\theta_i) \text{ is generated from the } k^{th} \text{ component.}\end{cases}$$

给定完整数据 $\{(\mathbf{x}_i^*,\theta_i,\mathbf{z}_i):\ i=1,2,\ldots,n\}$，对数似然为 s

$$\ell^c(\boldsymbol{\vartheta}|\boldsymbol{\theta},\mathbf{x}^*)=\sum_{i=1}^{n}\sum_{k=1}^{K}z_{ik}\Big[\log\pi_k+\log\mathcal{VM}(\theta_i|\mu_k+g(\mathbf{x}_i^{*\intercal}\boldsymbol{B}_k),\kappa_k)\Big]. \tag{20}$$

给定 $\boldsymbol{\vartheta}$的一个合适初始值，记为 $\boldsymbol{\vartheta}^{(0)}$，EM算法在期望（E-）步和最大化（M-）步这两个步骤之间迭代。在E步中，在第 r^{th}次迭代时，算法基于给定$\mathbf{z}$ 条件下$\mathbf{x}^*$, $\boldsymbol{\theta}$ 和 $\boldsymbol{\vartheta}^{(r-1)}$的条件分布，求得 $\ell^c(\boldsymbol{\vartheta}|\boldsymbol{\theta},\mathbf{x}^*)$的期望值，即

$$Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(r-1)})=\sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{E}[z_{ik}|\mathbf{x}_i,\theta_i,\boldsymbol{\vartheta}^{(r-1)}]\Big[\log\pi_k+\log\mathcal{VM}(\theta_i|\mu_k+g(\mathbf{x}_i^{*\intercal}\boldsymbol{B}_k),\kappa_k)\Big], \tag{21}$$

这可简化为计算 $\mathbb{E}[z_{ik}|\mathbf{x}_i^*, \theta_i, \boldsymbol{\vartheta}^{(r-1)}]$，其中 $i = 1, 2, \ldots, n$，即

$$\gamma_{ik}^{(r)} = \frac{\pi_k^{(r-1)} \mathcal{VM}(\theta_i|\mu_k^{(r-1)} + g(\mathbf{x}_i^{*\intercal}\boldsymbol{B}_k^{(r-1)}), \kappa_k^{(r-1)})}{\sum_{j=1}^K \pi_j^{(r-1)} \mathcal{VM}(\theta_i|\mu_j^{(r-1)} + g(\mathbf{x}_i^{*\intercal}\boldsymbol{B}_j^{(r-1)}), \kappa_j^{(r-1)})}. \tag{22}$$

在 M 步骤中，我们最大化 $Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(r-1)})$ 以更新 $\boldsymbol{\vartheta}^{(r-1)}$，在设定 $\mathbb{E}[z_{ik}|\mathbf{x}_i^*, \theta_i, \boldsymbol{\vartheta}^{(r-1)}] = \gamma_{ik}^{(r)}$，其中 $i = 1、2、\ldots\ldots、n$。

关于 μ_k 最大化 $Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(r-1)})$，对于 $k = 1、2、\ldots\ldots、K$，得到

$$\frac{\partial Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(r-1)})}{\partial \mu_k} = -\kappa_k \sum_{i=1}^n \gamma_{ik}^{(r)} \sin(\theta_i - \mu_k - g(\mathbf{x}_i^{*\intercal}\boldsymbol{B}_k)). \tag{23}$$

令 (23) 等于零并利用三角恒等式 (10) 求解 μ_k，可得到 $\mu_k^{(r-1)}$ 的更新公式，对于 $k = 1、2、\ldots\ldots、K$，为

$$\mu_k^{(r)} = \tan^{-1}\left[\frac{\sum_{i=1}^n \gamma_{ik}^{(r)} \sin(\theta_i - g(\mathbf{x}_i^{*\intercal}\boldsymbol{B}_k^{(r-1)}))}{\sum_{i=1}^n \gamma_{ik}^{(r)} \cos(\theta_i - g(\mathbf{x}\cdot *\intercal_i \boldsymbol{B}_k^{(r-1)}))}\right]. \tag{24}$$

关于 κ_k 最大化 $Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(r-1)})$，对于 $k = 1、2、\ldots\ldots、K$，得到

$$\frac{\partial Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(r-1)})}{\partial \kappa_k} = -n_k A(\kappa_k) + \sum_{i=1}^n \gamma_{ik}^{(r)} \cos(\theta_i - \mu_k - g(\mathbf{x}_i^\intercal \boldsymbol{\beta}_k)), \tag{25}$$

其中 $n_k = \gamma_{ik}^{(r)}$，对于 $k = 1、2、\ldots\ldots、K$。

令（25）等于零并求解 κ_k，即可得到 $\kappa_k^{(r-1)}$ 的更新方程，其中 $k = 1、2、\ldots\ldots、K$，如下所示

$$\kappa_k^{(r)} = A^{-1}\left(\frac{1}{n_k}\sum_{i=1}^n \gamma_{ik}^{(r)} \cos(\theta_i - \mu_k^{(r)} - g(\mathbf{x}_i^{*\intercal}\boldsymbol{B}_k^{(r-1)}))\right). \tag{26}$$

为更新回归系数 $\boldsymbol{B}_k^{(r-1)}$，其中 $k = 1、2、\ldots\ldots、K$，我们使用牛顿-拉夫森算法，相对于 $\boldsymbol{B}_k$ 迭代最大化 $Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(r-1)})$，其中 $k = 1、2、\ldots\ldots、K$。

$\boldsymbol{B}_k$ 的 NR更新规则，其中 $k = 1、2、\ldots\ldots、K$，如下所示

$$\boldsymbol{B}_k^{\text{new}} = \boldsymbol{B}_k^{\text{old}} - \mathbf{H}^{-1}(\boldsymbol{B}_k^{\text{old}})\boldsymbol{g}(\boldsymbol{B}_k^{\text{old}}), \tag{27}$$

其中

$$\boldsymbol{g}(\boldsymbol{B}_k) \equiv \frac{\partial Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(r-1)})}{\partial \boldsymbol{B}_k} = \mathbf{X}^{*\intercal}\dot{\boldsymbol{G}}_k \mathbf{W}_k^{(r)} \mathbf{u}_k, \tag{28}$$

$$\mathbf{H}(\boldsymbol{B}_k) \equiv \frac{\partial^2 Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(r-1)})}{\partial^2 \boldsymbol{B}_k} = -\mathbf{X}^{*\intercal}\dot{\boldsymbol{G}}_k^2 \mathbf{W}_k^{(r)} \boldsymbol{V}_k \mathbf{X}^* + \mathbf{X}^{*\intercal}\ddot{\boldsymbol{G}}_k \mathbf{W}_k^{(r)} \mathbf{U}_k \mathbf{X}^*, \tag{29}$$

其中 $\boldsymbol{W}_k^{(r)} = \text{diag}\{\gamma_{1k}^{(r)}, \gamma_{2k}^{(r)}, \ldots, \gamma_{nk}^{(r)}\}$，对于 $k = 1,\ 2,\ \ldots,\ K$。

对于每个 $k = 1$、2、.....、K，$\boldsymbol{G}_k$、$\mathbf{u}_k$、$\boldsymbol{V}_k$、$\boldsymbol{G}_k$和$\mathbf{U}_k$的定义方式与第2.3节类似。为了更新 $\boldsymbol{B}_k^{(r-1)}$，我们使用 $\boldsymbol{B}_k^{(r)}$，即在牛顿-拉夫森算法收敛时 $\boldsymbol{B}_k^{\text{old}}$ 的更新值，其中 $k = 1$、2、.....、K。我们重复上述 E 步和 M 步骤，直至收敛。上述估计程序总结于算法2中。

算法2 拟合圆形-线性回归混合模型

()1：分别为 μ_k、κ_k 和 $\boldsymbol{B}_k$适当选取初始值 $\mu_k^{(0)}$、$\kappa_k^{(0)}$ 和 $\boldsymbol{B}_k^{(0)}$，其中 $k = 1$、2、.....、K。2: **E步：** 在第 r^{th} 次迭代中，使用公式（22）计算后验概率 $\gamma_{ik}^{(r)}$，其中$i = 1$、2、.....、n以及 $k = 1$、2、.....、K。3: **M 步骤：** 在第 r^{th} 次迭代中，分别使用公式（24）、（26）和（27）更新 $\mu_k^{(r-1)}$、$\kappa_k^{(r-1)}$ 以及 $\boldsymbol{B}_k^{(r-1)}$，其中 $k = 1$、2、.....、K。4: 重复**E步** 和**M 步骤**，直至收敛。

2.6 建模与计算细节

2.6.1 基于模型的聚类规则

对于由模型 (18) 生成的每个观测（$\mathbf{x}_i, \theta_i$），其中 $i = 1, 2, \ldots, n$，令 $\mathbf{z}_i$ =($z_{i1}, z_{i2}, \ldots, z_{iK}$) 为簇成员指示向量，其中，若 ($\mathbf{x}_i, \theta_i$) 属于簇 k，则 $z_{ik} = 1$；否则为 $z_{ik} = 0$。

令 $\hat{\boldsymbol{\vartheta}}$ 为通过最大化对数似然 (19) 得到的 $\boldsymbol{\vartheta}$ 的最大似然估计。对于每个 $\mathbf{z}_i$，其中 $i = 1, 2, \ldots, n$，我们可以使用最大后验（ MAP ）估计：

$$\hat{z}_{ik} = \begin{cases} 1 & \text{if } \max_j\{\gamma_{ij}\} = \gamma_{ik} \\ 0 & \text{otherwise} \end{cases}$$

其中

$$\gamma_{ik} = \frac{\hat{\pi}_k \mathcal{VM}(\theta_i | \hat{\mu}_k + g(\mathbf{x}_i^{*\intercal} \hat{\boldsymbol{B}}_k), \hat{\kappa}_k)}{\sum_{j=1}^{K} \hat{\pi}_j \mathcal{VM}(\theta_i | \hat{\mu}_j + g(\mathbf{x}_i^{*\intercal} \hat{\boldsymbol{B}}_j), \hat{\kappa}_j)}, \tag{30}$$

是 ($\mathbf{x}_i, \theta_i$) 属于第 k^{th} 个成分的后验概率。

2.6.2 选择成分数

实际上，在估计模型（18）之前，我们需要选择成分数 K。对于大多数实际问题而言，这是一项艰巨的任务；若未能选择一个

合理的 K 取值，可能导致拟合模型不佳，表现为 K取值过小而拟合不足，或者K （Fruhwirth-Schnatter 等人 [2019]；第 7 章）所述的， K 取值过大导致的可辨识性问题。对于大多数现实问题，先验地即可选择，也就是说，在获得观测数据之前就确定。为此，人们采用信息准则等数据驱动方法来选择 K。本文提出使用贝叶斯信息准则（BIC）（Schwarz [1978]）

$$\mathrm{BIC}(K) = -2\ell(\hat{\boldsymbol{\vartheta}}) + \log(n) \times \mathrm{df}(K), \tag{31}$$

ˆ其中 $\boldsymbol{\vartheta}$ 是 $\boldsymbol{\vartheta}$ 的最大似然估计，df(K)是拟合的K–成分模型的估计参数个数。

最佳模型是使 BIC（K）（31）最小的模型，其中 $K \in \{1, 2, \ldots, K_{max}\}$取 K_{max} 所考虑的最大值。注意，对于 $K = 1$，圆形-线性回归混合模型（18）对应于圆形-线性回归模型。

2.6.3 计算平均方向

令 $S = \sum_{i=1}^{n} \sin(\theta_i - g(\mathbf{x}_i^{*\intercal}\boldsymbol{B}))$ 和 $C = \sum_{i=1}^{n} \cos(\theta_i - g(\mathbf{x}_i^{*\intercal}\boldsymbol{B}))$ 分别为式（11）中反正切项的分子和分母。根据Fisher[1995] 和 Jammalamadaka 和 Sengupta [2001], ， μ 在 $[0, 2\pi)$上的正确估计值应考虑 S 和 C的符号。因此，式（11）中的 ˆμ 定义为

$$\hat{\mu}^* = \begin{cases} \tan^{-1}(S/C) & \text{if } S > 0 \text{ and } C > 0, \\ \tan^{-1}(S/C) + \pi & \text{if } C < 0, \\ \tan^{-1}(S/C) + 2\pi & \text{if } S < 0 \text{ and } C > 0. \end{cases} \tag{32}$$

上述内容同样适用于 (24) 中的 $\mu_k^{(r)}$，其中 $k = 1, 2, \ldots, K$ 的定义方式与 (32) 类似，即 $\mu_k^{*(r)}$，其中 $k = 1, 2, \ldots, K$，且 $S = S_k^{(r)} = \sum_{i=1}^{n} \gamma_{ik}^{(r)} \sin(\theta_i - g(\mathbf{x}_i^{*\intercal}\boldsymbol{B}_k^{(r-1)}))$ 和 $C = C_k^{(r)} = \sum_{i=1}^{n} \gamma_{ik}^{(r)} \cos(\theta_i - g(\mathbf{x}_i^{*\intercal}\boldsymbol{B}_k^{(r-1)}))$，其中 $k = 1, 2, \ldots, K$。

2.6.4 初始化策略

关于混合建模，尤其是Shireman等人 [2017]的文献充分记载了EM算法对其初始值的依赖性（Biernacki等人[2003] 以及Fruhwirth-Schnatter等人 [2019]第2章）。这一问题受到了广泛关注，并由此产生了多种解决方案。本文采用

一种随机初始化策略。我们从 R 个不同的随机选取的参数值开始初始化算法，并选择产生最大对数似然值的解。在模拟和真实数据分析中，我们分别将 $R = 10$ 和 $R = 50$ 的值设为相应数值。根据模拟结果（见第3节），该策略效果良好。

3 模拟研究

本节通过大量蒙特卡洛模拟展示所提方法的性能。我们考察了多种情形，例如增加成分数量，以及改变成分之间的重叠程度。对于每种情形，我们分别生成样本量为 $n = 500$、1000 和 2000 的 1000 个样本。为便于说明，所有情形均使用一个线性协变量 $\mathbf{x}_2 = x$，它取自区间 $(-0.5, 0.5)$ 上的均匀分布；以及一个圆形协变量 $\mathbf{x}_1 = (\sin(x), \cos(x))$，它取自区间 $(\pi/3, 8\pi/3)$ 上的均匀分布。因此，模型（18）中的 $\mathbf{x}^* = (\mathbf{x}_1, \mathbf{x}_2)^\intercal$，下文将后者称为 MixCircReg 模型。

3.1 估计性能

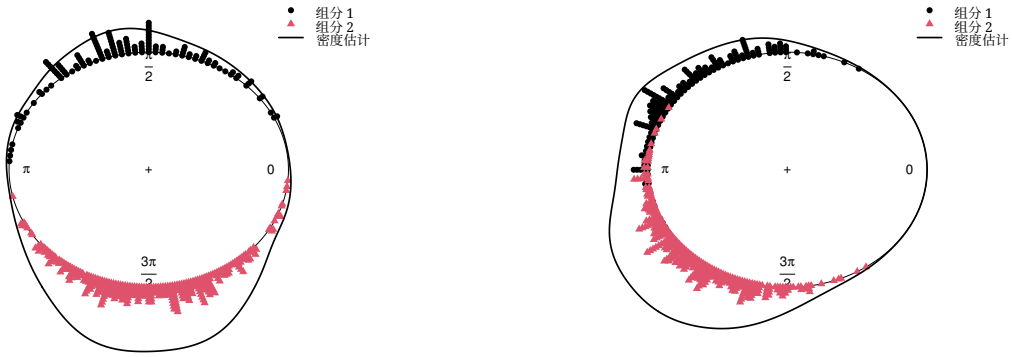
我们使用均方根误差（RMSE）来评估估计模型参数的性能。

3.1.1 场景 1： $K = 2$ 不重叠成分

在第一个示例中，我们展示所提方法在检测两个不重叠成分方面的性能。数据生成如下

$$\theta_i = \begin{cases} \mathcal{VM}(\mu_1 + g(\mathbf{x}_i^{*\top}\boldsymbol{B}_1), \kappa_1) \text{ with probability } \pi \\ \mathcal{VM}(\mu_2 + g(\mathbf{x}_i^{*\top}\boldsymbol{B}_2), \kappa_2) \text{ with probability } 1-\pi \end{cases} \quad \text{for } i = 1, 2, \ldots, n \tag{33}$$

其中 $\mu_1 = 1.8850$, $\mu_2 = 4.7124$, $\kappa_1 = 4$, $\kappa_2 = 6$, $\pi = 0.3$, $\boldsymbol{B}_1 = (\beta_{11}, \beta_{12}, \beta_{13})^\top$，其中 $\beta_{11} = 0.2$, $\beta_{12} = 0.1$ 和 $\beta_{13} = 0.3$，以及 $\boldsymbol{B}_2 = (\beta_{21}, \beta_{22}, \beta_{23})$，其中 $\beta_{21} = 0.1$, $\beta_{22} = 0.2$ 和$\beta_{23} = 0.2$。图 2a 展示了一个使用（33）生成的数据示例的圆形直方图（图中叠加了核密度估计），对应于 $n = 500$。从图 2a 可以清楚地看出，两个成分互不重叠。表 1 给出了拟合所得的结果



(a) 不重叠成分　　(b) 重叠成分

图 2：使用 (33) 生成的大小为 $n = 500$ 的典型样本的圆形直方图，其中 (a) 为不重叠情形，(b) 为重叠情形。

MixCircReg 模型用于所有生成的样本，样本量分别为 $n = 500$、1000 和 2000。可以看出，随着样本量增大，估计误差逐渐减小，这体现了所提出的最大似然估计量的收敛性质。

表 1：场景 1 和场景 2 在 1000 次重复试验中的参数估计值的均方根误差

参数	场景 1			场景 2		
	$n = 500$	$n = 1000$	$n = 2000$	$n = 500$	$n = 1000$	$n = 2000$
π	0.0206	0.0148	0.0105	0.0233	0.0184	0.0147
μ_1	0.0617	0.0461	0.0342	0.0803	0.0581	0.0393
μ_2	0.0331	0.0242	0.0171	0.0403	0.0298	0.0223
κ_1	0.5018	0.3410	0.2282	0.6699	0.4643	0.3764
κ_2	0.4622	0.3839	0.2527	0.5289	0.4344	0.2965
β_{11}	0.0507	0.0383	0.0282	0.0506	0.0417	0.0271
β_{12}	0.0541	0.0492	0.0456	0.0557	0.0482	0.0502
β_{13}	0.1344	0.1267	0.1221	0.1403	0.1232	0.1335
β_{21}	0.0541	0.0192	0.0141	0.0287	0.0246	0.0165
β_{22}	0.0489	0.0427	0.0336	0.0495	0.0445	0.0362
β_{23}	0.1267	0.1146	0.0902	0.1282	0.1231	0.0964

3.1.2 场景 2： $K = 2$ 重叠成分

现在，我们展示所提方法在检测两个重叠成分时的性能。数据使用式（33）生成，其中 $\mu_1 = 2.5133$ 和 $\mu_2 = 4.0841$。其余参数值与场景 1 相同。图 2b 展示了一个使用式（33）生成的数据示例的圆形直方图（图中叠加了核密度估计），对应于 $n = 500$。从图（2b）可以清楚地看出，这两个成分存在重叠。表2列出了将 MixCircReg 模型拟合到生成的全部 1000 个样本所得的结果，样本大小为 $n = 500$、1000 和 2000。结果的解读方式与场景 1 中所得结果类似。不过，与场景 1 相比，可以看出成分重叠时收敛速度较慢。

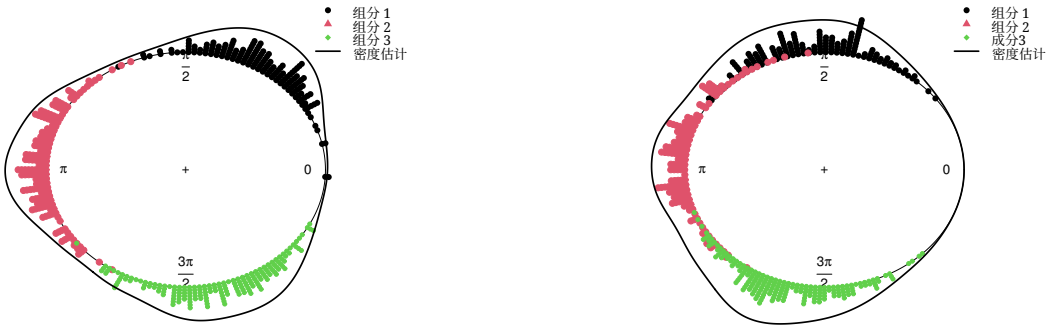
3.1.3 情景3： $K = 3$ 不重叠成分

接下来，我们展示所提方法在检测三个不重叠成分方面的性能。数据按如下方式生成

$$\theta_i \;= \begin{cases} \mathcal{VM}(\mu_1 + g(\mathbf{x}_i^{*\top}\boldsymbol{B}_1), \kappa_1) \;\text{ with probability }\; \pi_1 \\ \mathcal{VM}(\mu_2 + g(\mathbf{x}_i^{*\top}\boldsymbol{B}_2), \kappa_2) \;\text{ with probability }\; \pi_2 \\ \mathcal{VM}(\mu_3 + g(\mathbf{x}_i^{*\top}\boldsymbol{B}_3), \kappa_3) \;\text{ with probability }\; \pi_3 \end{cases} \tag{34}$$
$$\text{for } i = 1, 2, \ldots, n$$

其中 $\mu_1 = 1.0996$, $\mu_2 = 3.1416$, $\mu_3 = 5.0625$, $\kappa_1 = 8$, $\kappa_2 = 6$, $\kappa_3 = 8$, $\pi_1 = 0.33$, $\pi_2 = 0.33$, $\pi_3 = 0.34$, $\boldsymbol{B}_1 = (\beta_{11},\ \beta_{12},\ \beta_{13})^\top$，其中 $\beta_{11} = 0.085$, $\beta_{12} = 0.1$ 和 $\beta_{13} = 0.3$,$\boldsymbol{B}_2 = (\beta_{21},\ \beta_{22},\ \beta_{23})^\top$，其中 $\beta_{21} = 0.09$, $\beta_{22} = 0.1$ 和 $\beta_{23} = 0.2$ 以及 $\boldsymbol{B}_3 = (\beta_{31},\ \beta_{32},\ \beta_{33})$，其中 $\beta_{31} = 0.1$, $\beta_{32} = 0.1$ 和 $\beta_{33} = 0.1$。图 3a 展示了一个使用 (33) 生成的示例数据的圆形直方图（图中叠加了核密度估计），其中 $n = 500$。从图 3a 可以清楚看出，三个成分彼此分离良好，因此互不重叠。

表2给出了将 MixCircReg 模型拟合到生成的所有 1000 个样本（样本量分别为 $n = 500$、1000 和 2000）后得到的结果。结果表明，当成分数量增加到 $K = 3$ 时，所提最大似然估计量仍表现良好。



(a) 不重叠成分　　(b) 重叠成分

图3：使用（34）生成的一个典型样本的圆形直方图，样本量为 $n = 500$ ，对应于（a）不重叠和（b）重叠的情形。

3.1.4 情景4： $K = 3$ 个重叠成分

现在，我们展示所提方法在检测三个重叠成分时的表现。数据使用 (34) 生成，其中 $\mu_1 = 1.7279$、 $\mu_2 = 3.1416$ 和 $\mu_3 = 4.5553$。其余参数与情景3相同。图3b展示了使用 (33) 为 $n = 500$ 生成的一个数据集示例的圆形直方图（图上叠加了核密度估计）。从图 (2b) 可以清楚看出，这三个成分彼此重叠。

表2给出了对所有 1000 个生成的样本（样本量分别为 $n = 500$、1000 和 2000年）拟合 MixCircReg 模型所得到的结果。这些结果的解读方式与场景 2 中所得结果类似。

总之，所提出的估计程序在估计所有模型参数时表现良好且相近。然而，与其他参数估计值相比，κ 的模型参数估计值收敛速度较慢。所有考察的情景均如此。总体而言，所有考察情景所得结果表明，所提出的估计程序在各种条件下均表现令人满意。

表2：情景3和情景4在1000次重复试验中参数估计值的均方根误差

参数	情景3			情景4		
	$n = 500$	$n = 1000$	$n = 2000$	$n = 500$	$n = 1000$	$n = 2000$
π_1	0.0204	0.0149	0.0112	0.0228	0.0172	0.0135
π_2	0.0218	0.0149	0.0118	0.0281	0.0224	0.0146
π_3	0.0213	0.0154	0.0108	0.0279	0.0184	0.0120
μ_1	0.0404	0.0292	0.0219	0.0430	0.0307	0.0247
μ_2	0.0499	0.0361	0.0278	0.0508	0.0465	0.0315
μ_3	0.0419	0.0308	0.0216	0.0409	0.0342	0.0256
κ_1	0.9503	0.7129	0.5062	1.0596	0.8724	0.5553
κ_2	0.8418	0.5989	0.4083	1.5114	1.1159	0.8898
κ_3	1.0043	0.7479	0.4911	1.2476	0.8287	0.5711
β_{11}	0.0322	0.0235	0.0177	0.0362	0.0257	0.0196
β_{12}	0.0483	0.0447	0.0417	0.0528	0.0444	0.0429
β_{13}	0.1277	0.1166	0.1127	0.1316	0.1161	0.1113
β_{21}	0.0389	0.0287	0.0214	0.0333	0.0327	0.0284
β_{22}	0.0544	0.0509	0.0438	0.0590	0.0531	0.0516
β_{23}	0.1359	0.1336	0.1185	0.1366	0.1225	0.1276
β_{31}	0.0342	0.0240	0.0189	0.0329	0.0287	0.0199
β_{32}	0.0502	0.0461	0.0368	0.0478	0.0473	0.0357
β_{33}	0.1271	0.1190	0.0955	0.1336	0.1180	0.0938

3.2 聚类性能

本模拟研究旨在评估所提出模型将观测值分配到各个同质组中的能力。混合建模允许使用第 2.6.1 节所述的最大后验方法，将数据划分为若干簇。我们将使用误分类率（ClassErr）、ClassError $= \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(\hat{z}_{ik} \neq z_{ik})$ 和调整兰德指数（ARI；Hubert 和 Arabie [1985]）来衡量聚类过程的质量。ARI 的取值范围为 0 到 1。该指数考虑了数据点被随机分配到各簇的情况，因此在随机分类下其期望值为 0。我们计算 100 次重复试验的平均 ClassErr 和平均 ARI。ClassErr 和 ARI 使用 `classError()` 函数和

`adjustedRandIndex()`函数，分别来自 `R` 包 `mclust`。结果见表3。ClassErr 较小、ARI 较大，说明所提出的聚类过程表现良好。结果表明，对于所考虑的不重叠情形（场景 1 和 3）以及重叠情形（场景 2 和 4），这一结论均成立。不过，可以看出，与后者相比，该模型在前一种情形下的表现略好。当然，这是因为成分重叠时难以将它们分开。

表3：对 1000 次重复试验计算得到的误分类率（ClassErr）和调整兰德指数（ARI）的均值（及标准差）。

	场景 1		场景 2		情景3		情景4	
n	ClassErr	ARI	ClassErr	ARI	ClassErr	ARI	ClassErr	ARI
500	0.004	0.985	0.040	0.845	0.012	0.963	0.0533	0.850
	(0.003)	(0.011)	(0.0103)	(0.0389)	(0.005)	(0.015)	(0.0100)	(0.0273)
1000	0.003	0.987	0.040	0.843	0.012	0.965	0.0541	0.850
	(0.002)	(0.009)	(0.0066)	(0.0251)	(0.003)	(0.010)	(0.0084)	(0.0220)
2000	0.003	0.986	0.041	0.840	0.012	0.965	0.0537	0.850
	(0.002)	(0.006)	(0.0054)	(0.0209)	(0.003)	(0.008)	(0.0056)	(0.0150)

4 应用

在本节中，我们展示所提方法在此前第1节中介绍的风向数据集上的实际效用。

对于风能生产而言，风向是衡量风能潜力的最重要指标之一，即某一给定位置可利用的风能量（Rad 等 [2022]）。风电场（选址）的具体特征，如风速和气温，可能会影响风向。因此，了解风向对这些变量的依赖关系，有助于作出风电场选址和/或风力涡轮机布局这一重要决策，从而最大限度地提高风能生产的产出。在这一应用中，我们关注的是利用观测到的风速、气温以及一天中的时间（按24小时制计）来解释图1b中所绘风向的变化。

本应用所使用的风向数据是一个大型风向数据集的一部分，该数据集此前曾用于Rad 等 [2022], ，其中可以找到对该数据的详细描述。

表4给出了风数据的相关系数。任意两个线性变量之间的相关性使用通常的皮尔逊相关系数度量（Jammala- madaka 和 Lund [2006]）。另一方面，圆形变量风向和一天中的时间之间的相关性，使用 Jammalamadaka 和 Sengupta [2001]第8.2节中的圆形相关度量来衡量。而圆形变量风向与线性变量风速和气温之间的相关性，则使用 Jammalamadaka 和 Sengupta[2001]第8.5节中的相关度量来衡量。

如引言中所述，观测风向呈双峰分布（见图1a）。因此，我们拟合了 $K = 2$–成分模型

$$\mathcal{MVM}(\theta|\mathbf{X}^* = \mathbf{x}^*) = \pi_1\mathcal{VM}(\mu_1 + g(\mathbf{x}^*\boldsymbol{\beta}_1), \kappa_1) + \pi_2\mathcal{VM}(\mu_2 + g(\mathbf{x}^*\boldsymbol{\beta}_2), \kappa_2), \qquad (35)$$

其中 $\theta =$ 表示风向， $\phi =$ 表示一天中的小时， $\mathbf{x}^* = (\mathbf{x}_1, \mathbf{x}_2)^\intercal$，其中 $\mathbf{x}_1 = (\sin(\phi),\ \cos(\phi))^\intercal$，而 $\mathbf{x}_2 =$ (风速，气温)$^\intercal$。拟合模型的参数估计值见表5。我们还使用以下参数自助法计算了标准误和95%置信区间：

1. 对于给定的 x，我们从拟合模型 (35) 中抽取响应的相应值，记为 θ^*：$\pi_1\mathcal{VM}($ $\mu_1 + g(x\beta_1),\ \kappa_1) + \pi_2\mathcal{VM}(\mu_2 + g(x\beta_2),\ \kappa_2)$。重复这一抽样过程 n 次，得到一个自助样本 $\mathcal{S} = \{(x_i,\ \theta_i^*)\colon\ i = 1,\ 2,\ \ldots\ldots,\ n\}$。我们按照上述方式生成 B 个自助样本 $\mathcal{S}^{(1)},\ \mathcal{S}^{(2)},\ \ldots\ldots,\ \mathcal{S}^{(B)}$。

2. 接下来，我们在每个自助样本上拟合模型 (35)，从而生成 $k = 1,\ 2,\ \ldots\ldots,\ K$ 对应的 π_k、μ_k、κ_k 和 β_k 的抽样分布。

3. 最后，要计算 95% 置信区间，我们将抽样分布的 2.5^{th} 和 97.5^{th} 百分位数分别作为区间的下限和上限。我们设定 $B = 1000$。

从表 5 可以看出，大多数时间（约 57%）风都来自西北方向，估计风向约为 4.8（弧度），即 275。所有

估计的回归系数在统计上均显著，但风来自西北方向时的风速系数除外（第二个成分，估计风向约为 275）。

我们还拟合了圆回归模型（$K = 1$），得到的 BIC 为 2586.414，而拟合模型（35）的 BIC 为 1787。与后一个模型相比，这表明模型（35）对风数据的拟合效果最佳。

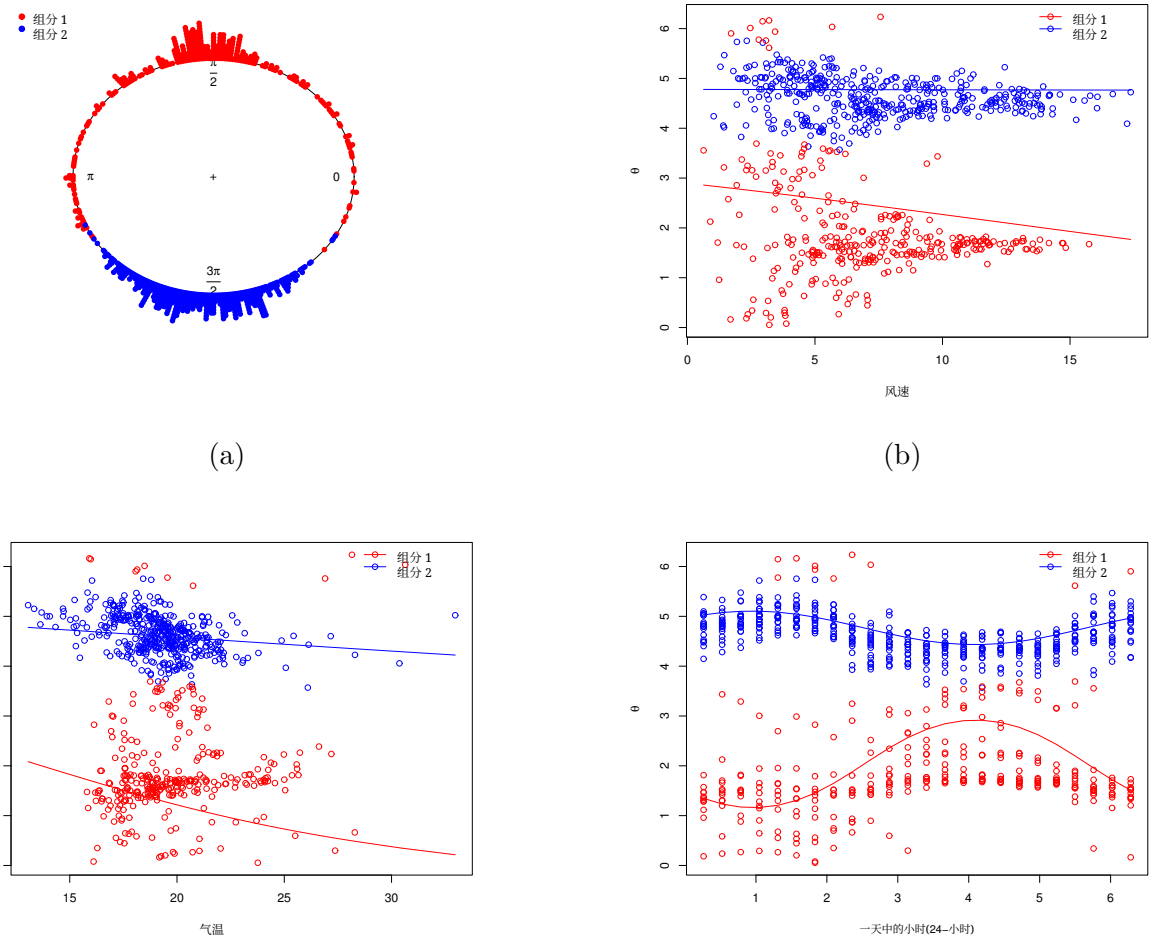
最后，为了展示拟合模型的聚类能力，我们可以使用基于模型的聚类规则，将每个观测分配到相应的风向类别。图 4a–4d 分别展示了圆形图，以及风向与风速、气温和一天中的小时的散点图。根据第 2.6.1 节所述的基于模型的聚类规则，按照各点所属的成分对其进行着色。散点图中还绘出了估计得到的各成分圆回归函数。

表 4：风数据的相关系数。

	Wind 方向-tion	风速	气温-ture	一天中的时间（小时）
风向	1	-	-	-
风速	0.0454*	1	-	-
气温	0.0390*	0.0505	1	-
Time of day（小时）	-0.4076*	0.0451*	0.1800*	1
*在 0.001 水平上显著				

表 5：基于风数据拟合模型 (35) 得到的参数估计值（含标准误（标准误差））、95% 置信区间和拟合统计量。标准误和置信区间通过所提出的参数自助法程序获得。

参数	估计值（标准误）	95% 自助法置信区间
π_1	0.4294 (0.0180)	(0.3934，0.4645)
μ_1	2.0388 (0.1382)	(1.7614, 2.3150)
μ_2	4.7690 (0.0497)	(4.6730, 4.8700)
κ_1	3.0825 (0.2516)	(2.6640, 3, 6167)
κ_2	13.4460 (1.0343)	(11.7314, 15.7270)
$\boldsymbol{\beta}_1$	-0.2695 (0.0367)	(-0.3448,-0.2046)
	-0.3838 (0.0376)	(-0.4619,-0.3164)
	-0.0679 (0.0417)	(-0.0921,-0.0456)
	-0.1897 (0.0842)	(-0.2702,-0.1270)
$\boldsymbol{\beta}_2$	0.0991 (0.0111)	(0.0775,0.1211)
	0.1382 (0.0115)	(0.1161,0.1604)
	-0.0009 (0.0152))	(-0.009,0.009)
	-0.0406 (0.0240)	(-0.0617,-0.0196)
$\ell(\hat{\vartheta})$	-851	
BIC	1787	



(a) (b)

(c) (d)

图4: （a）风向的圆形图，以及（b）—（d）风向分别对风速、气温和一天中的小时的散点图，其中每个点根据其所属成分，按照基于模型的聚类规则进行着色。红色点对应成分 1（东风向），蓝色点对应成分 2（西风向）。实线表示各成分拟合得到的回归函数。

5 结论

本文提出了一种圆回归模型混合模型，用于在数据呈多峰分布时，研究圆形响应对一组线性（或实值）和/或圆形协变量的依赖关系。为估计该模型，我们提出了使用期望最大化算法获得最大似然估计的方法。通过蒙特卡洛模拟展示了该模型和估计程序的性能。此外，为了证明所提模型的实用价值，我们将其应用于一个真实的风数据集，以研究风速、气温和一天中的时间对风向的影响。我们利用该拟合模型将观测风向分类到模型识别出的两个簇中，从而展示了该模型的优势。未来的研究可以考虑扩展所提模型，以处理圆形响应变量中的离群值（或非典型观测值）和偏度。

声明

利益冲突： 作者声明不存在与本文内容相关的任何利益冲突。

数据可用性：本文使用的数据和代码可通过以下链接获取：
https://github.com/Sphiwe-Skhosana/MixCircReg

参考文献

E. Batschelet。《style id='1'>《生物学中的圆统计》。生物学中的数学。Academic Press，1981。J. Bentley。使用冯·米塞斯分布与均匀分布混合物对圆形数据建模。硕士论文，西蒙菲莎大学，2006。C. Biernacki、G. Celeux 和 G. Govaert。为多元高斯混合模型的 EM 算法选择初始值，以获得最高似然。《计算统计与数据分析》，41(3-4)：561–575，2003。C.M. Bishop。《模式识别与机器学习》。Springer New York, NY, 2006。

G. E. P. Box 和 G. C. Tiao。针对若干离群值问题的贝叶斯方法。*Biometrika*，55(1)：119–129，1968。J. A. Carta、C. Bueno 和 P. Ramˊırez。利用冯·米塞斯分布混合物对方向性风速进行统计建模：案例研究。*Energy Conversion and Manage-ment*，49(5)：897–907，2008。A. R. Child、B. R. Howell 和 R. G. Houghton。泰晤士河口鳎（Solea solea (L.)）产卵的日周期性与时间。*ICES Journal of MarineScience*，48(3)：317–323，1991。A. P. Dempster、N. M. Laird 和 D. B. Rubin。通过 EM 算法从不完整数据中求最大似然。*Journal of the Royal Statistical Society: Series B (Methodological)*，39(1)：1–22，1977。W. S. DeSarbo 和 W. L. Cron。聚类线性回归的最大似然方法。《分类学杂志》，5：249–282，1988。T. D. Downs 和 K. V. Mardia。圆周回归。*Biometrika*，89(3)：683–698，2002。N. I. Fisher。圆形数据的统计分析。剑桥大学出版社，1995。N. I. Fisher 和 A. J. Lee。角响应的回归模型。*Biometrics*，第665–677页，1992。N. I. Fisher、T. Lewis 和 B. J. J. Embleton。球面数据的统计分析。剑桥大学出版社，1993。S. Fr¨uhwirth-Schnatter。有限混合与马尔可夫切换模型。Springer，纽约，2006。S. Fruhwirth-Schnatter、G. Celeux 和 C. P. Robert。混合分析手册。CRC Press，2019。S. M. Goldfeld 和 R. E. Quandt。通过切换回归估计结构突变。普林斯顿大学，1973。A. L. Gould。角变量回归技术。*Biometrics*，25(4)：683–700，1969。

L. Hubert 和 P. Arabie。比较划分。《分类学杂志》，2:193–218，1985年。S. Ingrassia、S. C. Minotti 和 A. Punzo。基于线性簇加权模型的基于模型的聚类。《计算统计与数据分析》，71:159–182，2014年。R. A. Jacobs、M. I. Jordan、S. J. Nowlan 和 G. E. Hinton。局部专家自适应混合模型。《神经计算》，3(1):79–87，1991年。S. R. Jammalamadaka 和 U. J. Lund。风向对臭氧浓度的影响：案例研究。《环境与生态统计》，13:287–298，2006年。S. R. Jammalamadaka 和 A. Sengupta。圆统计学专题，第5卷。世界科学出版社，2001年。S. R. Jammalamadaka 和 V. S. Vaidyanathan。通过冯·米塞斯分布有限混合对圆形数据进行聚类及其在风向数据中的应用。*SankhyaA*，第1–21页，2024年。R. A. Johnson 和 T. E. Wehrly。一些角度—线性分布及相关回归模型。《美国统计学会会刊》，73(363):602–606，1978年。L. N. Joseph、C. Elkin、T. G. Martin 和 H. P. Possingham。使用 n 混合模型对丰度建模：考虑生态机制的重要性。《生态学应用》，19(3):631–642，2009年。S. Kim 和 A. SenGupta。涉及圆形变量的回归：综述。载于《统计学及其应用》，第25–33页。Springer，新加坡，2018年。T. Kneib、A. Silbersdorff 和 B. S¨afken。反对均值——分布回归方法综述。《计量经济学与统计学》，26:99–123，2023年。S. J. Kon。股票收益模型——比较。《金融学杂志》，39(1):147–165，1984年。C. Ley 和 T. Verdebout。现代方向统计学。Chapman and Hall/CRC 出版社，2017。U. Lund。方向数据的最小圆距离回归。《应用统计学杂志》，26(6):723–733，1999年。K. V. Mardia 和 P. E. Jupp。方向统计学。约翰·威立父子出版公司，2000年。

I. Markovi´c 和 I. Petrovi´c。基于冯·米塞斯分布混合的仅方位跟踪。载于*2012*年*IEEE/RSJ*智能机器人与系统国际会议，第707–712页。IEEE，2012年。N. Masseran、A. M. Razali、K. Ibrahim 和 M. T. Latif。拟合冯·米塞斯分布混合，以对马来西亚半岛的风向数据建模。*EnergyConversion and Management*，72:94–102，2013年。G.J. McLachlan 和 D. Peel。有限混合模型。威利出版社，2000年。J. A. Mooney、P. J. Helms 和 I. T. Jolliffe。拟合冯·米塞斯分布混合：婴儿猝死综合征案例研究。《计算统计与数据分析》，41(3-4):505–513，2003年。S. Newcomb。关于组合观测以获得最佳结果的广义理论。《美国数学杂志》，第343–366页，1886年。K. Pearson。对进化数学理论的贡献。《伦敦皇家学会哲学汇刊*A*辑》，185:71–110，1894年。R. E. Quandt。一种估计切换回归的新方法。《美国统计学会会刊》，67(338):306–310，1972年。R. E. Quandt 和 J. B. Ramsey。正态分布混合与切换回归的估计。《美国统计学会会刊》，73(364):730–738，1978年。N. Rad、A. Bekker 和 M. Arashi。通过贝叶斯混合建模改进南非风能热点地区的风向预测。《科学报告》，12(1):11442，2022年。Y. Sarma 和 S. R. Jammalamadaka。圆周回归。《统计科学与数据分析》，34:109–128，1993年。M. P. A. Schneider 和 E. Scharfenaker。混合起来：使用有限混合模型研究收入分布的理由。《欧洲物理学杂志专题》，229 (9):1685–1704，2020年。G. Schwarz。模型维数的估计。《统计年鉴》，第461–464页，1978年。

E. Shireman、D. Steinley 和 M. J. Brusco。考察初始化策略对高斯混合建模性能的影响。行为研究方法，49:282– 293，2017。M. A. Stephens。方向数据技术。技术报告，统计学系，1969。W. Yao 和 S. Xiang。《混合模型：参数、半参数与新方向》。CRC Press，2024。C. B. Zeller、C. R. B. Cabral、V. H. Lachos 和 L. Benites。基于正态分布的尺度混合的删失数据有限混合回归模型。《数据分析与分类进展》，13:89–116，2019。