arXiv:2601.12425v1 [stat.ME] 18 Jan 2026

基于污染高斯分布的稳健半参数线性专家混合模型

PetersonMambondimumwe, Sphiwe B. Skhosana* 和 Najmeh Nakhaei Rad

统计系，比勒陀利亚大学，比勒陀利亚，0028，南非

*通讯作者： spiwe.skhosana@up.ac.za

2026年1月21日

摘要

半参数和非参数回归混合模型是一类非常实用且灵活的回归混合模型，其中部分或全部参数是协变量的非参数函数。然而，这些模型基于成分误差分布服从高斯分布的假设。因此，当存在离群值和重尾误差分布时，其估计结果会受到影响。本文提出半参数和非参数污染高斯回归混合模型，以便在存在轻度离群值时稳健地估计模型的参数项和/或非参数项。采用污染高斯误差分布的优点是，可以同时对观测进行基于模型的聚类和基于模型的离群值检测。我们提出两种算法：期望最大化（EM）型算法和期望条件最大化（ECM）型算法，分别对所提模型中的参数部分进行最大似然估计，对非参数部分进行局部似然核估计。我们通过大规模模拟研究考察了所提模型的稳健性，并利用真实数据展示了所提模型的实际应用价值。

关键词：期望最大化（*EM*）算法、非参数回归、回归混合模型、污染高斯分布

1 引言

当基础总体由数量通常未知的（设为 K 个）未观测子总体（也称为成分）构成时，线性回归模型的有限混合对于研究响应变量 y 与一组协变量 $\mathbf{x} = (x_1, x_2, \ldots\ldots, x_p)^\top$ 之间的关系非常有用。

设 Z 为一个分量指示变量，其服从离散分布 $P(Z = k) = \pi_k$，其中 $k = 1, 2, \ldots\ldots, K$，且每个 $\pi_k > 0$ 和 $\sum_{k=1}^{K} \pi_k = 1$。给定 $Z = k$ 和 $\mathbf{X} = \mathbf{x}$，y 与 $\mathbf{x}$ 之间的关系为

$$y = \mathbf{x}^\top \boldsymbol{\beta}_k + \epsilon_k, \text{ for } k = 1, 2, \ldots, K, \tag{1}$$

其中 $\mathbf{x} = (1, x_1, x_2, \ldots\ldots, x_p)^\top$，$\boldsymbol{\beta}_k = (\beta_{k0}, \beta_{k1}, \ldots\ldots, \beta_{kp})$ 为回归参数向量，ϵ_k 为误差项，通常假设其满足：(i) 与 $\mathbf{x}$ 独立；(ii) 服从均值为零、方差为 σ_k^2 的高斯分布。即 $\epsilon_k \sim \mathcal{N}(0, \sigma_k^2)$，其中 $\mathcal{N}(\mu, \sigma^2)$ 表示均值为 μ 、方差为 σ^2 的高斯分布。由于 Z 通常不可观测，仅给定 $\mathbf{X} = \mathbf{x}$ 时，y 被认为服从高斯分布混合，其条件密度函数为

$$f(y|\mathbf{X} = \mathbf{x}) = \sum_{k}^{K} \pi_k \mathcal{N}\{y|\mathbf{x}^\top \boldsymbol{\beta}_k, \sigma_k^2\}, \tag{2}$$

其中，权重 π_k' 是混合概率，用于表示各组件的相对规模。模型（2）是线性回归高斯混合模型（GMLRs）。

尽管灵活，线性回归混合（2）的适用性仍受限于混合比例为常数这一假设。为了增强模型的能力，Jacobs 等人 [1991] 提出了专家混合模型（MoE），允许混合比例成为一组协变量的函数。

令 $\mathbf{t} = (1, t_1, t_2, \ldots, t_q)^\top$ 为一组协变量，MoE 使用多项逻辑函数将混合比例建模为 $\mathbf{t}$的函数

$$\pi_k(\mathbf{t}|\boldsymbol{\gamma}) = P(Z = k|\mathbf{t}) = \frac{\exp(\mathbf{t}^\top \boldsymbol{\gamma}_k)}{\sum_{j=1}^{K} \exp(\mathbf{t}^\top \boldsymbol{\gamma}_j)}, \tag{3}$$

其中，$\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_{K-1})$，其中 $\boldsymbol{\gamma}_k = (\gamma_{k0}, \gamma_{k1}, \ldots, \gamma_{kq})$是第$k^{th}$ 个组件的系数向量，而 $\boldsymbol{\gamma}_K$ 是用于确保可识别性的零向量。

给定$\mathbf{x}$ 和$\mathbf{t}$，y 的条件混合密度为

$$f(y|\mathbf{X} = \mathbf{x}, \mathbf{T} = \mathbf{t}) = \sum_{k}^{K} \pi_k(\mathbf{t}|\boldsymbol{\gamma}) \mathcal{N}\{y|\mathbf{x}^\top \boldsymbol{\beta}_k, \sigma_k^2\}, \tag{4}$$

其中，对于所有$\mathbf{t}$，$\pi_k(\mathbf{t}|\boldsymbol{\gamma}) > 0$ 且 $\sum_{k=1}^{K} \pi_k(\mathbf{t}|\boldsymbol{\gamma}) = 1$。

模型 (4) 是高斯混合专家模型（GMoE），其中混合比例 $\pi_k(\mathbf{t}|\boldsymbol{\gamma})$ 称为门控函数，成分密度 $\mathcal{N}\{y|\mathbf{x}^\top\boldsymbol{\beta}_k, \sigma_k^2\}$ 称为专家。注意，协变量 $\mathbf{x}$ 和 $\mathbf{t}$ 可以相同。

MoE (4) 仍受限于门控函数 $\pi_k(\mathbf{t}|\boldsymbol{\gamma})$ 必须为多项逻辑函数这一假设。逻辑函数的单调形状可能并不合适（Young and Hunter [2010]）。为进一步提高 MoE 模型的灵活性，Young and Hunter [2010] 和 Huang and Yao [2012] 提出以非参数方式对门控函数建模

$$f(y|\mathbf{X} = \mathbf{x}, \mathbf{T} = \mathbf{t}) = \sum_{k}^{K} \pi_k(\mathbf{t})\mathcal{N}\{y|\mathbf{x}^\top\boldsymbol{\beta}_k, \sigma_k^2\}, \tag{5}$$

其中 $\pi_k(\mathbf{t})$ 是协变量 $\mathbf{t}$的未知光滑函数，因此是非参数的。

考虑随机样本 $\{(\mathbf{x}_i, \mathbf{t}_i, y_i) : i = 1, 2, \ldots, n\}$，Young and Hunter [2010] 提出将非参数门控函数建模为 $\pi_k(\mathbf{t}_i) = \mathbb{E}\{Z_{ik}|\mathbf{t}_i\}$，其中 $Z_{ik} = 1$ 表示数据点由第 i^{th} 个专家生成，否则为 0。Young and Hunter [2010]采用局部多项式回归方法，通过将 Z_{ik} 作为拟合所得 GMLRs (2) 的输出，获得门控函数的非参数估计。Huang and Yao [2012]建议采用局部似然方法（Tibshirani and Hastie [1987]）对非参数门控函数进行完全非参数建模。上述非参数估计方法无法处理高维协变量。为解决这一局限，Xiang and Yao [2020]提出使用单指标模型 $\pi_k(\mathbf{t}_i^\top\boldsymbol{\gamma}_k)$ 对非参数门控函数建模，其中 $\boldsymbol{\gamma}_k$ 是单指标参数向量。作为替代方案，Xue and Yao [2022]提出使用神经网络对非参数门控函数建模。

请注意，迄今讨论的模型均假设每个成分都服从高斯分布。然而在实践中，数据可能受到离群值污染，这可能会对模型估计产生显著的负面影响Mazza 和 Punzo [2020]。 [Mazza 和 Punzo, 2020] 区分了轻度离群值和严重离群值。作者将轻度离群值定义为在给定组件内与回归关系偏离不显著的点；然而，这些点会导致分量分布的尾部过重，无法用高斯分布充分建模。另一方面，严重离群值是指与任何回归组件都相距甚远的点。本文关注轻度离群值。关于这两类离群值的区别，更多详情请参见

可参阅 Mazza 和 Punzo [2020] 及其中的参考文献。对于高斯回归混合模型，最常用的离群值防护方法是用重尾成分替换每个高斯分量 Mazza 和 Punzo [2020]。Yao 等人 [2014]假设每个成分都服从 t分布，提出了稳健回归混合模型（下文简称 t-MR）。Song 等人[2014] 也提出了类似的稳健方法，但假设每个成分都服从拉普拉斯分布（下文简称 L-MR）。最近，Mazza 和 Punzo [2020] 使用污染高斯分布对回归模型混合中各成分的分布进行建模（下文简称 CG-MLR）。Zeller 等人 [2016] 提出了一种更通用的方法，通过使用偏斜高斯尺度混合分布对每个成分建模，同时考虑非对称性（下文简称 SMSG-MR）。然而，Yao 等人 [2014], 、Song 等人 [2014], 和 Mazza 和 Punzo [2020] 提出的稳健回归模型混合都假设混合比例为常数。Chamroukhi [2016],、Nguyen 和 McLachlan [2016], 以及 Mirfarah 等人 [2021] 分别扩展了 Yao 等人 [2014], 、Song 等人 [2014], 和 Mazza 和 Punzo [2020], 的模型，提出了稳健专家混合（MoE）模型，并将混合比例建模为一组协变量的函数。下文分别将前三种模型称为 t-MoE、L-MoE 和 CG-MoE。请注意，CG-MoE 是 Mirfarah 等人 [2021]提出的模型的一个特例。不过，该模型对于本文的研究目标而言过于通用。还需注意，t-MoE、L-MoE 和 CG-MoE 均使用多项逻辑函数对门控函数或混合比例建模。因此，其实际应用受限于这一参数假设。增强稳健专家线性混合模型灵活性的一种显而易见且简单的方法，是对门控函数进行非参数建模。然而，目前这方面的研究尚未受到多少关注。唯一的例外是 Ge 等人 [2024], ，他们提出扩展 t-MoLE 模型，假设门控函数是协变量的非参数函数。下文将该模型称为 S-t-MoLE。作者证明，当成分分布的尾部比高斯分布更重时，S-t-MoLE 模型具有稳健性和实际效用。

本文提出一种稳健的半参数线性专家混合模型，假设每个组件的分布都服从污染高斯分布。下文将所得模型称为 S-CG-MoLE。与 S-t-MoLE 相比，本文提出的 S-CG-MoLE 采用一种有依据的概率方法，同时进行观测聚类——

和轻度离群值检测。将 S-CG-MoLE 模型拟合到数据后，可先将每个观测分配到某个组件或聚类，随后再将其分类为典型或非典型（离群值）观测（参见 Mazza 和 Punzo [2020]）。本文的贡献如下：

1. 我们提出一种稳健且灵活的回归模型混合，假设每个组件均服从污染高斯分布，且混合比例函数是协变量的非参数函数。

2. 我们介绍一种改进的期望条件最大化（ECM）算法，用于同时进行参数和非参数最大似然估计。

本文其余部分安排如下。第2节回顾污染高斯线性专家混合（CG-MoLE）及其基于期望条件最大化（ECM）算法的估计过程（总结于算法1）。随后，我们将CG-MoLE扩展为所提出的混合比例可变的半参数CG-MoLE模型，并介绍其估计过程（总结于算法2）。本节最后分别讨论如何利用所提出的模型进行模型选择和基于模型的聚类。第3节通过模拟研究，评估所提出模型在各种情景下的稳健性。第4节将所提出的模型应用于一个真实数据集。最后，第5节总结全文并提出未来研究方向。

2 方法论

2.1 污染高斯线性专家混合模型

本节定义 CG-MoLE 模型，并介绍如何使用最大似然法对其进行估计。

模型定义

CG-MoLE 模型是作为Mirfarah 等人 [2021], 提出的模型的一种特殊情况而得到的，是高斯混合线性专家模型（GMoLE）的扩展。该 K–组件 CG-MoLE 模型的形式为

$$f(y|\mathbf{X}=\mathbf{x},\mathbf{T}=\mathbf{t})=\sum_{k=1}^{K}\pi_k(\mathbf{t}|\boldsymbol{\gamma})\mathcal{CN}\{y|\mathbf{x};\boldsymbol{\theta}_k\}, \tag{6}$$

其中

$$\mathcal{CN}\{y|\mathbf{x};\boldsymbol{\theta}_k\} = \alpha_k\mathcal{N}\{y|\mathbf{x}^\top\boldsymbol{\beta}_k,\sigma_k^2\} + (1-\alpha_k)\mathcal{N}\{y|\mathbf{x}^\top\boldsymbol{\beta}_k,\eta_k\sigma_k^2\} \tag{7}$$

是 k^{th} 组件的污染高斯分布密度函数，而 $\boldsymbol{\theta}_k = (\alpha_k, \boldsymbol{\beta}_k, \eta_k, \sigma_k^2)$是 k^{th} 污染高斯专家的参数向量。参数 $\alpha_k \in (0,1)$表示典型（未受污染）数据点的比例，$\eta_k > 1$ 则表示污染程度。后一个参数衡量离群值导致方差膨胀的程度。

令 $\boldsymbol{\Theta} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_K; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_K)$为CG-MoLE 模型的参数向量。CG-MoLE 模型参数可通过最大似然估计得到。接下来我们将对此进行讨论。

模型估计

考虑来自CG-MoLE模型（6）的随机样本 $\{(\mathbf{x}_i, \mathbf{t}_i, y_i) : i = 1, 2, \ldots, n\}$。相应的观测对数似然函数为

$$\ell\{\boldsymbol{\Theta}\} = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k(\mathbf{t}_i|\boldsymbol{\gamma})[\alpha_k\mathcal{N}\{y|\mathbf{x}^\top\boldsymbol{\beta}_k,\sigma_k^2\} + (1-\alpha_k)\mathcal{N}\{y|\mathbf{x}^\top\boldsymbol{\beta}_k,\eta_k\sigma_k^2\}]. \tag{8}$$

$\boldsymbol{\Theta}$的最大似然估计为

$$\hat{\boldsymbol{\Theta}}_{\text{MLE}} = \max_{\boldsymbol{\Theta}} \ell\{\boldsymbol{\Theta}\}. \tag{9}$$

由于存在缺失数据，表达式（9）没有闭式解。更具体地说，缺失数据有两个来源。第一个缺失数据来源于我们不知道哪个组件生成了哪个数据点。第二个缺失数据来源于我们不知道组件中的某个数据点是离群值还是非离群值。因此，k^{th} 的最大似然估计 $\boldsymbol{\Theta}$ 是通过使用期望-条件最大化（ECM）算法最大化观测对数似然函数 $\ell\{\boldsymbol{\Theta}\}$获得的 Meng 和 Rubin [1993]。ECM 是传统期望最大化（EM）算法Dempster 等人[1977] 的扩展，用于在存在缺失数据时最大化观测对数似然。ECM 在收敛前会迭代执行三个步骤：一个期望（E）步和两个条件最大化（CM）步。两个 CM 步骤源于划分 $\boldsymbol{\Theta} = \{\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2\}$，其中 $\boldsymbol{\Theta}_1 = \{\alpha_k, \boldsymbol{\beta}_k, \sigma_k^2, \boldsymbol{\gamma}_k\}_{k=1}^K$ 以及 $\boldsymbol{\Theta}_2 = \{\eta_k\}_{k=1}^K$。E步计算

期望完全数据对数似然，而 CM 步骤对其进行最大化。完全数据为 $\{(\mathbf{x}_i,\ \mathbf{r}_i,\ y_i,\ \mathbf{z}_i,\ \mathbf{v}_i):\ i=1,\ 2、...、\ n\}$，其中 $\mathbf{z}_i=(z_{i1}、...、\ z_{iK})$；当 i^{th} 数据点来自 k^{th} 组件时，$z_{ik}=1$，否则为 0；并且 $\mathbf{v}_i=(v_{i1}、...、\mathbf{v}_{iK})$；当 i^{th} 组件中的 k^{th} 数据点是正常数据时，$v_{ik}=1$，否则为 $v_{ik}=0$。基于完全数据的对数似然为

$$\ell_c\{\mathbf{\Theta}\}=\ell_{1c}\{\boldsymbol{\gamma}\}+\ell_{2c}\{\boldsymbol{\alpha}\}+\ell_{3c}\{\boldsymbol{\vartheta}\},\tag{10}$$

其中

$$\begin{aligned}
\ell_{1c}\{\boldsymbol{\gamma}\} &= \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\ln\pi_k(\mathbf{t}_i|\boldsymbol{\gamma}),\\
\ell_{2c}\{\boldsymbol{\alpha}\} &= \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}[v_{ik}\ln\alpha_k+(1-v_{ik})\ln(1-\alpha_k)],\\
\ell_{3c}\{\boldsymbol{\vartheta}\} &= -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\left[z_{ik}\ln\sigma_k^2+z_{ik}(1-v_{ik})\ln\eta_k+z_{ik}\left(1+\frac{1-v_{ik}}{\eta_k}\right)\frac{(y_i-\mathbf{x}_i^\top\boldsymbol{\beta}_k)^2}{\sigma_k^2}\right],
\end{aligned}\tag{11}$$

with $\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\ldots,\alpha_K)$ and $\boldsymbol{\vartheta}=(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\ldots,\boldsymbol{\beta}_K;\sigma_1^2,\sigma_2^2,\ldots,\sigma_K^2;\eta_1,\eta_2,\ldots,\eta_K)$.

E步 在 E步中，在 ECM 算法的第 r^{th} 次迭代中，我们计算 $\ell_c\{\mathbf{\Theta}\}$的条件期望，记为 $Q\{\mathbf{\Theta}|\mathbf{\Theta}^{(r-1)}\}$。这归结为分别计算 $\mathbb{E}\{z_{ik}|y_i,\mathbf{x}_i,\ \mathbf{t}_i\}$和 $\mathbb{E}\{v_{ik}|y_i,\mathbf{x}_i,\ \mathbf{t}_i,\ \mathbf{z}_i\}$，其中 $i=1,\ 2,$ ....., n 以及 $k=1,\ 2,$ ....., K，如下所示：

$$\mathbb{E}\{z_{ik}|y_i,\mathbf{x}_i,\mathbf{t}_i\}\equiv z_{ik}^{(r)} = \frac{\pi_k(\mathbf{t}_i|\boldsymbol{\gamma}^{(r-1)})\mathcal{CN}\{y_i|\mathbf{x}_i;\boldsymbol{\theta}_k^{(r-1)}\}}{\sum_{\ell=1}^{K}\pi_\ell(\mathbf{t}_i|\boldsymbol{\gamma}^{(r-1)})\mathcal{CN}\{y_i|\mathbf{x}_i;\boldsymbol{\theta}_\ell^{(r-1)}\}},\tag{12}$$

$$\mathbb{E}\{v_{ik}|y_i,\mathbf{x}_i,\mathbf{t}_i,\mathbf{z}_i\}\equiv v_{ik}^{(r)} = \frac{\alpha_k^{(r-1)}\mathcal{N}\{y_i|\mathbf{x}_i^\top\boldsymbol{\beta}_k^{(r-1)},\sigma_k^{2(r-1)}\}}{\mathcal{CN}\{y_i|\mathbf{x}_i;\boldsymbol{\theta}_k^{(r-1)}\}}.\tag{13}$$

公式 (12) 可解释为 i^{th} 数据点属于 k^{th} 组件的后验概率，而公式 (13) 可解释为i^{th} 数据点在组件 k中不是离群值的后验概率。将 (10) 中的 z_{ik} 替换为 $z_{ik}^{(r)}$，并将 v_{ik} 替换为 $v_{ik}^{(r)}$，可得

$$Q\{\mathbf{\Theta}|\mathbf{\Theta}^{(r-1)}\}=Q\{\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(r-1)}\}+Q\{\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(r-1)}\}+Q\{\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(r-1)}\},\tag{14}$$

其中

$$
\begin{aligned}
Q\{\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(r-1)}\} &= \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}^{(r)} \ln \pi_k(\mathbf{t}_i|\boldsymbol{\gamma}), \\
Q\{\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(r-1)}\} &= \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}^{(r)} [v_{ik}^{(r)} \ln \alpha_k + (1-v_{ik}^{(r)}) \ln(1-\alpha_k)], \\
Q\{\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(r-1)}\} &= -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} \left[z_{ik}^{(r)} \ln \sigma_k^2 + z_{ik}^{(r)}(1-v_{ik}^{(r)}) \ln \eta_k + z_{ik}^{(r)} \left(1 + \frac{1-v_{ik}^{(r)}}{\eta_k}\right) \frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_k)^2}{\sigma_k^2} \right].
\end{aligned} \tag{15}
$$

CM步 1 在 ECM 算法的第 r^{th} 次迭代中，我们通过在 $\boldsymbol{\Theta}_2^{(r-1)}$ 固定时最大化 $Q\{\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(r-1)}\}$，来计算 $\boldsymbol{\Theta}_1^{(r)}$。具体而言，分别针对 $k = 1, 2, \ldots, K$，最大化 $Q\{\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(r-1)}\}$ 关于 α_k 的函数，以及针对 $k = 1,2, \ldots, K$，最大化 $Q\{\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(r-1)}\}$ 关于 $\boldsymbol{\beta}_k$ 和 σ_k^2 的函数，可得

$$
\alpha_k^{(r)} = \frac{1}{n_k^{(r)}} \sum_{i=1}^{n} z_{ik}^{(r)} v_{ik}^{(r)}, \tag{16}
$$

$$
\boldsymbol{\beta}_k^{(r)} = (\mathbf{X}^\top \mathbf{W}_k^{(r)} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}_k^{(r)} \mathbf{y}, \tag{17}
$$

$$
\sigma_k^{2(r)} = \frac{1}{n_k^{(r)}} \sum_{i=1}^{n} w_{ik}^{(r)} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_k^{(r)})^2, \tag{18}
$$

其中，$n_k^{(r)} = \sum_{i=1}^{n} z_{ik}^{(r)}$，$\mathbf{X} = (\mathbf{x}_1^\top, \mathbf{x}_2^\top, \ldots, \mathbf{x}_n^\top)^\top$ 是 $n \times (p+1)$ 设计矩阵，$\mathbf{y} = (y_1, y_2, \ldots, y_n)$，而 $\mathbf{W}_k^{(r)} = \text{diag}\{w_{1k}^{(r)}, w_{2k}^{(r)}, \ldots, w_{nk}^{(r)}\}$ 是一个对角矩阵，其中 $w_{ik}^{(r)} = z_{ik}^{(r)} \left(v_{ik}^{(r)} + \frac{1-v_{ik}^{(r)}}{\eta_k^{(r-1)}} \right)$，对于 $i = 1, 2, \ldots, n$ 和 $k = 1,2, \ldots, K$。

关于 $Q\{\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(r-1)}\}$ 相对于 $\boldsymbol{\gamma}_k$，对于 $k = 1$，2、……、K，无法得到闭式解。通常使用牛顿-拉夫森算法进行数值求解Bishop [2006]。

给定初始值 $\boldsymbol{\gamma}_k^{(0)}$，在牛顿-拉夫森算法的第 s^{th} 次迭代中，我们按如下方式更新 $\boldsymbol{\gamma}_k^{(s)}$

$$
\boldsymbol{\gamma}_k^{(s)} = \boldsymbol{\gamma}_k^{(s-1)} - \mathbf{H}^{-1}[\boldsymbol{\gamma}_k^{(s-1)}] \mathbf{g}[\boldsymbol{\gamma}_k^{(s-1)}], \tag{19}
$$

其中，$\mathbf{g}[\boldsymbol{\gamma}_k^{(s-1)}] = \frac{\partial Q\{\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(r-1)}\}}{\partial \boldsymbol{\gamma}_k}$ 和 $\mathbf{H}[\boldsymbol{\gamma}_k^{(s-1)}] = \frac{\partial^2 Q\{\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(r-1)}\}}{\partial^2 \boldsymbol{\gamma}_k}$ 分别是 $Q\{\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(r-1)}\}$的梯度向量和海森矩阵。$\boldsymbol{\gamma}_k^{(r)}$ 的值取为在式（19）中牛顿-拉夫森算法收敛时 $\boldsymbol{\gamma}_k$的值，其中 $k = 1$，2，……，K。

CM步 2 在第二个 CM 步中，在 ECM 算法的第 r^{th} 次迭代时，我们通过最大化 $Q\{\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(r-1)}\}$ 来计算 $\boldsymbol{\Theta}_2^{(r)}$，其中 $\boldsymbol{\Theta}_1$ 固定为在 CM步 1 中获得的 $\boldsymbol{\Theta}_1^{(r)}$。具体而言，我们最大化

$$-\frac{1}{2}\sum_{i=1}^{n} z_{ik}^{(r)}(1-v_{ik}^{(r)})\ln\eta_k - \frac{1}{2}\sum_{i=1}^{n} z_{ik}^{(r)}\frac{(1-v_{ik}^{(r)})}{\eta_k}\frac{(y_i-\mathbf{x}_i^{\top}\boldsymbol{\beta}_k^{(r)})^2}{\sigma_k^{2(r)}} \tag{20}$$

关于 η_k，在约束条件 $\eta_k > 1$ 下，对 $k = 1$，2、.....、K。

上述 ECM 算法总结于算法1中。

算法1 用于拟合 CG-MoLE 的 ECM 算法

令 $r = 0$ 表示当前迭代次数，$\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k, \sigma_k^2, \alpha_k, \eta_k)$ 和 $\boldsymbol{\psi}_k = (\boldsymbol{\gamma}_k, \boldsymbol{\beta}_k, \sigma_k^2, \alpha_k, \eta_k)$。该算法在 E步和两个 CM 步骤之间交替执行，直至收敛。

E步：使用公式 (12) 和 (13)，分别计算 $i = 1$、2、.....、n 和 $k = 1$、2、.....、K 对应的后验概率 $z_{ik}^{(r)}$ 和 $v_{ik}^{(r)}$。

CM步 1：更新 $k = 1$、2、.....、K 对应的参数 α_k、$\boldsymbol{\beta}_k$、σ_k^2，以及 $k = 1$、2、.....、$K-1$ 对应的参数 $\boldsymbol{\gamma}_k$。

- 给定 $\eta_{(r-1)\,k}$，利用式（16）、（17）和（18）计算 $\alpha_{(r)\,k}$、$\boldsymbol{\beta}_{(r)\,k}$ 和 $\sigma^2{}_{(r)\,k}$。
- 通过对 $\boldsymbol{\gamma}_{(s)\,k}$ 反复应用牛顿-拉夫森更新（19），直至收敛，从而计算 $\boldsymbol{\gamma}_{(r)\,k}$。

CM步 2：给定 $\boldsymbol{\beta}_k^{(r)}$ 和 $\sigma_k^{2(r)}$，通过在约束$\eta_k > 1$下最大化（34）来更新 η_k，其中 $k = 1$，2，...，K。这可以使用 R 编程语言中的 `optimize` 函数来实现。

重复上述 E 步和 CM 步骤，直至收敛。

2.2 污染高斯分布线性专家的半参数混合模型

尽管 CG-MoLE 模型 (6) 提供了一个在存在轻度离群值时仍具有稳健性的建模框架，但其适用性受到混合

比例函数 $\pi_k(\mathbf{t}|\gamma_k)$ 为多项逻辑函数这一假设的限制。为了增强 CG-MoLE 模型的能力，我们提出以非参数方式对混合比例函数建模。本节通过引入半参数 CG-MoLE（S-CG-MoLE）模型，对 CG-MoLE 模型 (6) 进行扩展。该模型假设混合比例函数是协变量 $\mathbf{t}$ 的非参数函数，而其他模型参数则与 CG-MoLE 模型中的定义相同。为估计非参数混合比例，我们开发了一种基于局部似然核的估计量，并使用 ECM 算法求得。已知核估计量会受到维数灾难的影响（参见 Nagler 和 Czado [2016]）。为避免这一问题，我们仅考虑单个协变量 t。不过请注意，所提出的方法可扩展至多变量 $\mathbf{t}$的情形。未来研究将探讨实现这一目标的方向。

2.2.1 模型定义

令 Z 为一个分量指示变量，其服从离散分布 $P(Z = k|T = t) = \pi_k(t)$，其中 $k = 1, 2, \ldots K$，且对所有 t，每个 $\pi_k(t) > 0$，而 $\sum_{i=1}^{K} \pi_k(t) = 1$。给定 $Z = k$，$\mathbf{X} = \mathbf{x}$ 和 $T = t$，假设 y 服从污染高斯分布，其密度函数为 $\mathcal{CN}\{y|\mathbf{x}; \boldsymbol{\theta}_k\}$ (7)。仅给定 $\mathbf{X} = \mathbf{x}$ 和 $T = t$ 时， y 服从污染高斯分布混合，其密度函数为

$$f(y|\mathbf{X} = \mathbf{x}, T = t) = \sum_{k=1}^{K} \pi_k(t)\mathcal{CN}\{y|\mathbf{x}; \boldsymbol{\theta}_k\}, \tag{21}$$

其中， $\pi_k(t)$ 被假设为协变量 t 的光滑未知函数。模型 (21) 即为所提出的半参数 CG-MoLE（S-CG-MoLE）模型。

模型 (21) 的估计与模型 (6) 的估计类似；唯一的区别在于，混合比例函数采用基于核的估计量进行非参数估计，而非使用 Newton-Raphson 算法进行参数估计。下一节将讨论模型 (21) 的估计。

2.2.2 模型估计

考虑来自 S-CG-MoLE 模型的随机样本 $\{(\mathbf{x}_i, t_i, y_i): \ i = 1, 2, \ldots, n\}$。相应的观测对数似然函数为

$$\ell\{\boldsymbol{\Psi}\} = \sum_{i=1}^{n} \ln \sum_{k=1}^{K} \pi_k(t_i)[\alpha_k \mathcal{N}\{y_i|\mathbf{x}_i^\top \boldsymbol{\beta}_k, \sigma_k^2\} + (1 - \alpha_k)\mathcal{N}\{y_i|\mathbf{x}_i^\top \boldsymbol{\beta}_k, \eta_k\sigma_k^2\}], \tag{22}$$

其中 $\boldsymbol{\Psi}=(\boldsymbol{\pi},\boldsymbol{\theta})$、 $\boldsymbol{\pi}=(\boldsymbol{\pi}_1(t)\ \boldsymbol{\pi}_2(t)\ldots,\boldsymbol{\pi}_K(t))$，其中 $\boldsymbol{\pi}_k=\{\pi_k(t_i)\colon\ i=1,2,\ldots,n\}$，且 $\boldsymbol{\theta}=(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\ldots,\boldsymbol{\theta}_K)$，其中 $\boldsymbol{\theta}_k=(\alpha_k,\boldsymbol{\beta}_k,\sigma_k^2,\eta_k)$。

与 CG-MoLE 模型 (6) 一样，S-CG-MoLE模型的估计通过使用 ECM 算法最大化观测数据对数似然 $\ell\{\boldsymbol{\Psi}\}$ 来进行。对于 S-CG-MoLE模型，两个 CM 步骤来自划分 $\boldsymbol{\Psi}=\{\boldsymbol{\Psi}_1,\boldsymbol{\Psi}_2\}$，其中 $\boldsymbol{\Psi}_1=\{\alpha_k,\boldsymbol{\beta}_k,\sigma_k^2,\boldsymbol{\pi}_k\}_{k=1}^K$和 $\boldsymbol{\Psi}_2=\{\eta_k\}_{k=1}^K$。相应的完全数据对数似然为

$$\ell_c\{\boldsymbol{\Psi}\}=\ell_{1c}\{\boldsymbol{\pi}\}+\ell_{2c}\{\boldsymbol{\alpha}\}+\ell_{3c}\{\boldsymbol{\vartheta}\},\tag{23}$$

其中 $\ell_{1c}\{\boldsymbol{\pi}\}=\sum_{i=1}^n\sum_{k=1}^K z_{ik}\ln\pi_k(t_i)$。完全数据对数似然 $\ell_{2c}\{\boldsymbol{\alpha}\}$ 和$\ell_{3c}\{\boldsymbol{\vartheta}\}$的定义见 (10)。

E步 在 E步中，在 ECM 算法的第 r^{th} 次迭代中，我们计算数据点 i^{th} 属于组件 k^{th} 的后验概率，记为 $z_{ik}^{(r)}$；以及数据点 i^{th} 在组件 k中不是离群值的后验概率，记为 $v_{ik}^{(r)}$，分别为

$$z_{ik}^{(r)} = \frac{\pi_k^{(r-1)}(t_i)\mathcal{CN}\{y_i|\mathbf{x}_i;\boldsymbol{\theta}_k^{(r-1)}\}}{\sum_{\ell=1}^K\pi_\ell^{(r-1)}(t_i)\mathcal{CN}\{y_i|\mathbf{x}_i;\boldsymbol{\theta}_\ell^{(r-1)}\}},\tag{24}$$

$$v_{ik}^{(r)} = \frac{\alpha_k^{(r-1)}\mathcal{N}\{y_i|\mathbf{x}_i^\top\boldsymbol{\beta}_k^{(r-1)},\sigma_k^{2(r-1)}\}}{\mathcal{CN}\{y_i|\mathbf{x}_i;\boldsymbol{\theta}_k^{(r-1)}\}}.\tag{25}$$

在（23）中用 $z_{ik}^{(r)}$ 替换 z_{ik}，并用 $v_{ik}^{(r)}$ 替换 v_{ik}，可得

$$Q\{\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(r-1)}\}=Q\{\boldsymbol{\pi}|\boldsymbol{\pi}^{(r-1)}\}+Q\{\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(r-1)}\}+Q\{\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(r-1)}\},\tag{26}$$

其中

$$Q\{\boldsymbol{\pi}|\boldsymbol{\pi}^{(r-1)}\} = \sum_{i=1}^n\sum_{k=1}^K z_{ik}^{(r)}\ln\pi_k(t_i).$$

函数 $Q\{\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(r-1)}\}$ 和 $Q\{\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(r-1)}\}$ 的定义与（14）类似。

CM步 1 在第一个 CM步中，在 ECM 算法的第 r^{th} 次迭代中，我们通过在将 $\boldsymbol{\Psi}_2$ 固定为 $\boldsymbol{\Psi}_2^{(r-1)}$ 时最大化 $Q\{\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(r-1)}\}$ 来计算 $\boldsymbol{\Psi}_1^{(r)}$。具体而言，对于 $k=1$，2，..，K，通过关于 α_k 最大化 $Q\{\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(r-1)}\}$，并通过关于 $\boldsymbol{\beta}_k$ 最大化 $Q\{\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(r-1)}\}$

分别对于 $k = 1, 2, \ldots, K$，我们得到 σ_k^2 和

$$\alpha_k^{(r)} = \frac{1}{n_k^{(r)}} \sum_{i=1}^{n} z_{ik}^{(r)} v_{ik}^{(r)}, \tag{27}$$

$$\boldsymbol{\beta}_k^{(r)} = (\mathbf{X}^\top \mathbf{W}_k^{(r)} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}_k^{(r)} \mathbf{y}, \tag{28}$$

$$\sigma_k^{2(r)} = \frac{1}{n_k^{(r)}} \sum_{i=1}^{n} w_{ik}^{(r)} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_k^{(r)})^2. \tag{29}$$

方程 (27)–(29) 的定义方式与方程 (16)–(18) 类似。

为了计算 $\pi_k(t)$ 的非参数估计，针对 $k = 1, 2, \ldots, K$，我们将使用基于核的局部线性估计量Fan and Gijbels [1996]。我们选择局部线性估计量，是因为它具有良好的性质，Fan [1992]对此进行了详细讨论。首先注意到，混合比例函数可定义为

$$\pi_k(t_i) = \mathbb{E}\{Z_{ik} | T_i = t_i\} \tag{30}$$

因此，$\pi_k(t_i)$ 是一个回归函数，其响应变量为 Z_{ik}，协变量为 t_i。由于我们无法观测到成分指示变量 Z_{ik}，因此将 $z_{ik}^{(r)}$（即在 ECM 算法的 E 步中得到的估计值）用作响应变量。

对于局部点 $u \in \mathcal{U}$ 邻域内的 t，其中 $\mathcal{U}$ 是协变量 t 定义域上的局部点集合，$\pi_k(t)$ 可在局部近似为一个线性函数

$$\pi_k(t) \approx \pi_k(u) + \pi_k'(u)[t - u] \equiv a_k + b_k(t - u). \tag{31}$$

由此得到以下局部加权最小二乘准则

$$-\frac{1}{2} \sum_{i=1}^{n} \left[z_{ik}^{(r)} - a_k - b_k(t_i - u) \right]^2 \mathcal{K}_h(t_i - u), \tag{32}$$

其中，$\mathcal{K}_h(\cdot) = \mathcal{K}(\cdot/h)/h$ 是一个重缩放核函数，用于为 u 附近的点赋予权重；h 是一个平滑参数，用于控制 u周围局部邻域的大小。常数 $1/2$ 是为计算方便而引入的。

对 a_k 和 b_k最大化（32），得到局部线性估计量

$$\pi_k^{(r)}(u) \equiv a_k^{(r)} = \frac{\sum_{i=1}^{n} [s_2(u) - s_1(u)(t_i - u)] \mathcal{K}_h(t_i - u) z_{ik}^{(r)}}{s_2(u) s_0(u) - s_1^2(u)}, \tag{33}$$

其中，$s_j(u) = \sum_{i=1}^{n} \mathcal{K}_h(t_i - u)(t_i - u)^j$，对于 $j = 0, 1, 2$。

通过对所有 $u \in \mathcal{U}$ 使用（33）获得非参数曲线估计。为获得$\pi_k^{(r)}(t_i)$，对于 $t_i \notin \mathcal{U}$，我们采用线性插值。需要注意的是，如果样本量不大，可以利用协变量 t 的观测值作为局部点。

CM步 2 在第二个 CM步中，在 ECM 算法的第 r^{th} 次迭代中，我们通过最大化 $Q\{\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(r-1)}\}$ 来计算 $\boldsymbol{\Psi}_2^{(r)}$，其中 $\boldsymbol{\Psi}_1$ 固定为在 CM步 1 中获得的 $\boldsymbol{\Psi}_1^{(r)}$。具体而言，我们最大化

$$-\frac{1}{2}\sum_{i=1}^{n} z_{ik}^{(r)}(1-v_{ik}^{(r)})\ln\eta_k - \frac{1}{2}\sum_{i=1}^{n} z_{ik}^{(r)}\frac{(1-v_{ik}^{(r)})}{\eta_k}\frac{(y_i-\mathbf{x}_i^\top\boldsymbol{\beta}_k^{(r)})^2}{\sigma_k^{2(r)}}, \tag{34}$$

关于 η_k，在约束 $\eta_k > 1$下，对于 $k = 1, 2, ..., K$。

上述 ECM 算法总结于算法2。

算法2 用于拟合 S-CG-MoLE 的 ECM 算法

令r 表示当前迭代的索引，给定上一轮迭代的估计值$\boldsymbol{\theta}_k^{(r-1)} = (\boldsymbol{\beta}_k^{(r-1)}, \sigma_k^{2(r-1)}, \alpha_k^{(r-1)}, \eta_k^{(r-1)})$和$\boldsymbol{\psi}_k^{(r-1)} = (\boldsymbol{\gamma}_k^{(r-1)}, \boldsymbol{\beta}_k^{(r-1)}, \sigma_k^{2(r-1)}, \alpha_k^{(r-1)}, \eta_k^{(r-1)})$，该算法在 E步和两个 CM步之间交替进行，直至收敛。

E步：使用公式（24）和（25）分别计算 $i = 1, 2, ..., n$ 和$k = 1, 2, ..., K$对应的后验概率 $z_{ik}^{(r)}$ 和 $v_{ik}^{(r)}$。

CM步 1：更新参数 α_k, $\boldsymbol{\beta}_k$, σ_k^2，针对 $k = 1, 2, \ldots, K$，以及非参数混合比例函数 $\pi_k(t_i)$，针对 $i = 1, 2, \ldots, n$和 $k =$**1**, 2, $\ldots, K-1$。

- **参数估计：** 给定 $\eta_{(r-1)\ k}$，使用 (16)、(17) 和 (18) 计算 $\alpha_{(r)\ k}$、$\boldsymbol{\beta}_{(r)\ k}$ 和 $\sigma^2{}_{(r)\ k}$
- **非参数估计：** 对于协变量 t 定义域内的每个局部点 $u \in \mathcal{U}$，使用 (33) 计算 $\pi_{(r)\ k}(u)$，并通过线性插值获得 $\pi_{(r)\ k}(t_i)$，针对 $t_i \notin \mathcal{U}$。

CM步 2：给定 $\boldsymbol{\beta}_k^{(r)}$ 和 $\sigma_k^{2(r)}$，通过在约束条件$\eta_k > 1$下最大化 (34) 来更新 η_k，针对 $k = 1, 2, \ldots, K$。这可以使用 `R` 编程语言中的 `optimize` 函数来完成。

重复上述 E 步和 CM 步骤，直至收敛。

2.3 模型选择

对于所提出的 S-CG-MoLE (21)，模型选择涉及选择 K（专家的数量）、 h（平滑参数）以及 $\mathcal{K}(\cdot)$（核函数）。为选择 K，我们采用定义如下的贝叶斯信息准则（BIC）：

$$\mathrm{BIC}(K) = -2\hat{\ell} + df \times \log(n), \tag{35}$$

其中 ℓ 是拟合模型的最大对数似然值，而 df 是根据拟合模型中参数的数量计算的自由度。需要注意的是，所提出的 S-CG-MoLE 模型同时包含参数部分和非参数部分。因此，我们定义 $df = df_1 + df_2$，其中 $df_1 = 3K + K(p+1)$ 和 $df_2 = K \times$EDF 分别表示模型参数部分和非参数部分的自由度。EDF 是一维平滑未知函数的有效自由度，例如$\pi_k(\cdot)$ Greenand Silverman [1994]。遵循文献中的常用方法（Huanget al. [2013], Xiangand Yao [2020], Skhosana et al. [2024]），我们将 EDF 定义为

$$\mathrm{EDF} = \tau_K h^{-1} |\mathcal{T}| \{\mathcal{K}(0) - \frac{1}{2}\int \mathcal{K}^2(u)du\}, \tag{36}$$

其中 $\mathcal{T}$ 是 t 的支撑集，并且

$$\tau_K = \frac{\mathcal{K}(0) - \frac{1}{2}\int \mathcal{K}^2(u)du}{\int \{\mathcal{K}(u) - \frac{1}{2}\mathcal{K} * \mathcal{K}(u)\}^2 du}. \tag{37}$$

带宽 h 采用与Huangand Yao[2012]中的相同交叉验证程序进行选择。

最后，核函数的选择通常并不那么重要，因为它不会影响局部线性估计量的收敛性 WuandZhang [2006]。特别是在非参数回归混合模型中，高斯核

$$\mathcal{K}(u) = \left(\sqrt{2\pi}\right)^{-1} \exp(-u^2/2) \tag{38}$$

迄今为止，它是实践中使用最广泛的核函数之一Xiang and Yao [2018]。

2.4 基于模型的聚类

对于任何污染高斯混合模型，都可以分两个阶段对观测值进行分类。第一阶段，我们确定其簇归属。随后，在第二

阶段，我们确定该观测值是否为该簇中的离群值。对于 S-CG-MoLE 模型（21），第一阶段使用最大后验算子（MAP）

$$\text{MAP}(\hat{z}_{ik}) = \begin{cases} 1 & \text{if } \max_j \hat{z}_{ij} \text{ occurs in component } k = j \\ 0 & \text{otherwise,} \end{cases} \tag{39}$$

其中，$\hat{}z_{ik}$ 是 ECM 算法 2 收敛时 $z_{ik}^{(r)}$ 的取值。根据（39），观测值（$\mathbf{x}_i$，t_i，y_i）将被分配到聚类 j。随后进入第二阶段：如果满足 $\hat{}v_{ik}^{(r)} < 0.5$，则将分配到聚类 j 的观测值（$\mathbf{x}_i$，t_i，y_i）标记为离群值，其中 $\hat{}v_{ik}$ 是 ECM 算法 2 收敛时 $v_{ik}^{(r)}$ 的取值。

3 模拟研究

本节将开展广泛的数值研究，以评估所提方法的性能。我们将所提出的 S-CGMoLE 模型与

1. 线性回归高斯混合模型（GMLRs）（2）；2. 污染线性回归高斯混合模型（CGMLRs）Mazza 和 Punzo[2020]；3. 线性专家高斯混合模型（GMoLE）（4）；4. 污染线性专家高斯混合模型（CGMoLE）（6）；5. 采用非参数门控函数的半参数线性专家高斯混合模型（S-GMoLE）（5）。

模拟使用 `R` 编程语言R 核心团队 [2023]。GMLRs 和 GMoLE 分别使用 `regmixEM`函数和 `hmeEM` 函数进行估计，这些函数来自 `R`软件包 `mixtools` Benaglia 等人 [2010]。目前没有用于拟合 CGMLRs、CGMoLE 和 S-GMoLE 的现成软件包。因此，为了估计这些模型，我们在 `R` 中编写了通过 EM 和/或 ECM 估计程序进行最大似然估计的代码。为评估各方法的估计精度，我们采用以下指标：

$$\text{MSE}(\hat{\boldsymbol{\pi}}(x)) \quad = \quad \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\left(\pi_k(x_i) - \hat{\pi}_k(x_i)\right)^2,$$

其中，ˆ$\pi_k(x)$ 和 $\pi_k(x)$ 分别是混合比例函数$\pi_k(x)$ 的估计值和真实值。

$$\text{MSE}(\hat{\theta}_k) \quad = \quad (\hat{\theta}_k - \theta_k)^2 \text{ and } \text{BIAS}(\hat{\theta}_k) = \hat{\theta}_k - \theta_k,$$

ˆ其中， θ_k 和 θ_k 分别是参数 θ_k的估计值和真实值。

3.1 数值研究

本研究的数据由 $K = 2$ 组件半参数线性回归混合模型生成。对于 $i = 1$，2，.....， n，每个 y_i都按如下方式独立生成

$$y_i = \begin{cases} x_i + \epsilon_{i1} & \text{if } z_i = 1 \\ 4 + x_i + \epsilon_{i2} & \text{if } z_i = 2, \end{cases}$$

其中， z_i 是 i^{th} 数据点的成分指示变量，其分布为 $P(z_i = 1) = \pi_1(x_i) = 0.1 + 0.8\sin(\pi x_i)$; x_i 由区间(0，1)上的均匀分布生成， ϵ_{ik} 是 k^{th} 线性回归模型的误差项。我们假设误差项 ϵ_{i1} 和 ϵ_{i2} 服从与 ϵ相同的分布。我们考虑 ϵ的以下分布：

1. $\epsilon \sim \mathcal{N}\{0, 1\}$;
2. $\epsilon \sim 0.95\mathcal{N}\{0, \sigma^2\} + 0.05\mathcal{N}\{0, \eta\sigma^2\}$，其中 $\sigma = 1$和 $\eta = 20$;
3. $\epsilon \sim$ t分布(ν)，具有 $\nu = 3$ 自由度;
4. $\epsilon \sim \mathcal{N}\{0, 1\}$，其中 y值中有10%被随机替换为噪声数据点，这些噪声数据点由区间（-15，15）上的均匀分布生成。

第一种情景用于评估所提出方法在数据未受离群值污染时的性能。第二种和第三种情景用于评估所提出方法在数据中存在离群值、因而数据服从重尾分布时的性能。第二种情景还用于评估所提出的算法2在拟合S-CG-MoLE模型时的性能。在每种情景下，我们生成样本量分别为 $n = 200$、500和1000的100个样本。表1-3分别给出了 $n = 200$、500和1000时参数估计值的均方误差和偏差。此外，表4给出了拟合得到的混合比例函数和参数的均方误差的平均值和标准差。

表 1–4 的主要结论如下：

- 对于情景（a），正如预期，由于分量误差分布服从高斯分布，SGMoE、GMoE 和 GMLRs 是整体表现最好的模型。不过，当 $n = 1000$ 较大时，这些模型的表现与本文提出的 SCGMoE 模型几乎相同。这表明，即使数据中不存在离群值，本文提出的 SCGMoE 模型也能表现良好。

- 对于情景（b），正如预期，所提出的 SCGMoE、CGMoE 和 CGMLR 整体表现最佳，因为分量误差分布服从污染高斯分布。另一方面，即使将样本量增加到 $n = 1000$，SGMoE、GMoE 和 GMLRs 这些模型的表现也没有改善。这表明，当假设分量分布为高斯分布时，离群值会对参数估计值产生负面影响。

- 对于情景（c），SCGMoE、SGMoE、CGMoE 和 CGMLR 这些模型整体表现最佳，且各模型的表现相近。另一方面，GMoE 和 GMLRs 表现最差，而且随着样本量增加到 $n = 1000$，其表现并未改善。值得注意的是，尽管 SGMoE 模型假设分量误差分布为高斯分布，但其表现依然良好。这凸显了在存在离群值时，正确设定混合比例模型的重要性。若模型设定错误，由此产生的偏差会导致分量特定参数估计值出现偏差。可参见 GMLRs 和 GMoE 的结果，这两个模型对混合比例的函数形式设定有误。

- 对于情景（d），CGMLR、CGMoE 和本文提出的 S-CGMoE 是总体表现最佳的模型。另一方面，即使将样本量增加到$n = 1000$，GMLRs、GMoE 和 SGMoE 这些模型的表现也没有改善。

总之，这项模拟研究表明，当数据具有重尾特征（情景（b）和（c））以及数据中受到含噪数据点污染时（情景（d）），采用重尾分量误差分布十分重要。此外，研究还表明，当数据中存在离群值且假设分量误差分布为高斯分布时，混合比例函数设定错误会如何影响模型参数估计值。

表 1： $n = 200$ 拟合模型参数的均方误差（偏差）（数值乘以 100）

	GMLRs	CGMLR	GMoE	CG-MoE	S-G-MoE	S-CG-MoE
			情景（a）			
β_{10}	0.010	0.001	0.001	0.001	0.003	0.030
	(-0.77)	(-0.33)	(0.053)	(-0.10)	(-0.51)	(-1.652)
β_{11}	0.080	0.090	0.021	0.026	0.020	0.034
	(2.78)	(3.05)	(1.445)	(1.61)	(1.40)	(1.86)
β_{20}	0.022	0.012	0.060	0.001	0.004	0.021
	(1.476)	(1.08)	(2.353)	(0.18)	(0.66)	(-1.45)
β_{21}	0.002	0.002	0.001	0.001	0.001	0.001
	(0.45)	(0.487)	(-0.200)	(0.12)	(-0.19)	(0.29)
σ_1	0.004	0.500	0.002	0.781	0.010	0.770
	(-0.63)	(-7.08)	(-0.403)	(-8.84)	(-0.90)	(-8.75)
σ_2	0.100	0.970	0.120	0.930	0.050	0.730
	(-3.18)	(-9.83)	(-3.48)	(-9.63)	(-2.26)	(-8.56)
			情景（b）			
β_{10}	9.41	0.07	97.51	0.020	0.520	0.000
	(-30.68)	(2.69)	(-98.75)	(1.560)	(7.240)	(-0.560)
β_{11}	90.96	0.53	422.38	0.420	1.23	0.35
	(95.37)	(-7.28)	(205.52)	(-6.470)	(-11.08)	(-5.91)
β_{20}	0.11	0.02	1.11	0.00	0.42	0.03
	(3.31)	(1.42)	(-10.53)	(0.06)	(-6.51)	(-1.75)
β_{21}	672.26	0.01	449.02	0.00	1.11	0.00
	(-259.28)	(-0.74)	(-211.90)	(-0.33)	(-10.52)	(-0.32)
σ_1	48.24	0.12	52.10	0.14	28.88	0.25
	(69.46)	(-3.47)	(72.18)	(-3.74)	(53.74)	(-5.00)
σ_2	114.16	0.27	72.07	0.24	45.90	0.150
	(106.85)	(-5.20)	(84.89)	(-4.900)	(67.75)	(-3.820)
			情景（c）			
β_{10}	1.764	0.062	206.998	0.571	1.046	1.621
	(-13.281)	(-2.484)	(-143.874)	(-7.556)	(-10.226)	(-12.733)
β_{11}	10.74	0.102	976.696	1.602	2.165	1.878
	(32.773)	(3.202)	(312.521)	(12.658)	(14.713)	(13.704)
β_{20}	20.549	0.192	8.726	0.007	0.018	0.217
	(45.331)	(4.377)	(-29.54)	(0.859)	(1.328)	(-4.655)
β_{21}	268.752	0.782	1325.462	0.043	0.555	0.072
	(-163.937)	(-8.841)	(-364.069)	(-2.066)	(-7.45)	(-2.686)
			情景（d）			
β_{10}	76.277	0.019	475.948	0.001	4.805	0.017
	(-87.337)	(1.381)	(-218.162)	(-0.200)	(21.92)	(-1.323)
β_{11}	619.26	0.007	1854.437	0.016	0.047	0.005
	(248.849)	(-0.823)	(430.632)	(1.276)	(2.178)	(0.687)
β_{20}	16.874	0.044	9.442	0.005	19.202	0.001
	(-41.078)	(2.095)	(-30.728)	(0.734)	(-43.82)	(-0.119)
β_{21}	575.425	0.185	1869.823	0.114	1.976	0.171
	(-239.88)	(-4.303)	(-432.414)	(-3.374)	(-14.058)	(-4.14)
σ_1	665.253	0.084	287.047	0.135	182.512	0.21
	(257.925)	(-2.895)	(169.425)	(-3.678)	(135.097)	(-4.587)
σ_2	325.515	0.694	209.874	0.396	96.371	0.297
	(180.42)	(-8.329)	(144.87)	(-6.29)	(98.169)	(-5.453)

表 2: $n = 500$ 的拟合模型参数的均方误差（偏差）（数值已乘以 100）

	GMLRs	CGMLR	GMoE	CG-MoE	S-G-MoE	S-CG-MoE
			情景（a）			
β_{10}	0.005	0.014	0.022	0.019	0.010	0.001
	(0.727)	(1.161)	(1.488)	(1.389)	(1.005)	(0.293)
β_{11}	0.001	0.001	0.024	0.016	0.014	0.017
	(-0.090)	(-0.133)	(-1.54)	(-1.272)	(-1.171)	(-1.308)
β_{20}	0.007	0.004	0.020	0.001	0.004	0.009
	(0.864)	(0.67)	(1.41)	(0.393)	(0.644)	(-0.940)
β_{21}	0.001	0.002	0.002	0.001	0.001	0.001
	(0.331)	(0.41)	(-0.456)	(-0.259)	(-0.245)	(-0.208)
σ_1	0.002	0.21	0.002	0.324	0.004	0.334
	(-0.468)	(-4.59)	(-0.409)	(-5.695)	(-0.649)	(-5.781)
σ_2	0.012	0.256	0.014	0.28	0.005	0.205
	(-1.099)	(-5.06)	(-1.18)	(-5.27)	(-0.697)	(-4.53)
			情景（b）			
β_{10}	1.07	0.00	39.49	0.08	0.47	0.01
	(-10.34)	(-0.12)	(-62.84)	(2.86)	(6.88)	(0.95)
β_{11}	128.1	0.02	413.07	0.17	0.15	0.1
	(113.18)	(1.27)	(203.24)	(-4.15)	(-3.91)	(-3.13)
β_{20}	3.74	0.00	3.53	0.08	1.45	0.00
	(-19.35)	(0.7)	(-18.78)	(2.75)	(-12.04)	(0.64)
β_{21}	591.53	0.00	631.8	0.09	0.43	0.05
	(-243.21)	(0.33)	(-251.36)	(-3.07)	(-6.54)	(-2.25)
σ_1	176.21	0.08	81.68	0.07	28.1	0.15
	(132.75)	(-2.75)	(90.38)	(-2.55)	(53.01)	(-3.81)
σ_2	310.49	0.05	102.07	0.05	73.76	0.01
	(176.21)	(-2.25)	(101.03)	(-2.29)	(85.88)	(-0.92)
			情景（c）			
β_{10}	42.363	0.00	510.69	0.096	0.00	0.474
	(-65.087)	(0.168)	(-225.984)	(-3.101)	(-0.113)	(-6.888)
β_{11}	290.835	0.232	2172.475	0.28	0.519	0.555
	(170.539)	(4.812)	(466.098)	(5.291)	(7.204)	(7.45)
β_{20}	0.621	0.087	0.53	0.06	0.747	0.354
	(-7.878)	(-2.951)	(-7.282)	(-2.458)	(-8.645)	(-5.949)
β_{21}	168.954	0.097	2307.075	0.059	0.001	0.14
	(-129.982)	(-3.115)	(-480.32)	(2.422)	(0.312)	(3.744)
			情景（d）			
β_{10}	2.644	0.117	155.56	0.027	11.144	0.063
	(-16.262)	(-3.418)	(-124.724)	(-1.644)	(33.383)	(-2.515)
β_{11}	296.406	0.506	1230.275	0.083	0.138	0.088
	(172.164)	(7.114)	(350.753)	(2.877)	(-3.719)	(2.959)
β_{20}	70.995	0.015	17.198	0.004	29.726	0.024
	(-84.258)	(-1.228)	(-41.471)	(-0.624)	(-54.522)	(-1.561)
β_{21}	315.975	0.029	2534.418	0.003	0.141	0.002
	(-177.757)	(1.701)	(-503.43)	(-0.538)	(3.756)	(-0.408)
σ_1	1199.269	0.03	362.874	0.038	265.506	0.068
	(346.305)	(-1.744)	(190.492)	(-1.954)	(162.944)	(-2.614)
σ_2	483.682	0.245	293.23	0.117	76.835	0.084
	(219.928)	(-4.949)	(171.24)	(-3.422)	(87.656)	(-2.901)

表 3: $n = 1000$ 拟合模型参数的均方误差（偏差）（数值乘以 100）

	GMLRs	CGMLR	GMoE	CG-MoE	S-G-MoE	S-CG-MoE
			情景（a）			
β_{10}	0.00	0.00	0.001	0.001	0.00	0.004
	(0.33)	(0.63)	(-0.245)	(0.253)	(0.010)	(-0.657)
β_{11}	0.004	0.01	0.00	0.001	0.00	0.00
	(-0.653)	(-0.760)	(0.54)	(-0.078)	(0.008)	(-0.006)
β_{20}	0.006	0.00	0.00	0.001	0.003	0.006
	(0.806)	(0.614)	(0.59)	(0.299)	(0.547)	(-0.755)
β_{21}	0.023	0.02	0.01	0.01	0.01	0.01
	(-1.502)	(-1.543)	(-0.88)	(-1.165)	(-1.15)	(-1.161)
σ_1	0.001	0.10	0.00	0.12	0.001	0.13
	(-0.177)	(-3.190)	(-0.14)	(-3.420)	(-0.15)	(-3.56)
σ_2	0.00	0.19	0.00	0.220	0.002	0.17
	(-0.59)	(-4.38)	(-0.64)	(-4.692)	(-0.499)	(-4.13)
			情景（b）			
β_{10}	3.73	0.00	115.29	0.00	0.11	0.01
	(-19.31)	(-0.18)	(-107.37)	(-0.22)	(3.38)	(-0.97)
β_{11}	123.99	0.00	650.2	0.00	0.04	0.00
	(111.35)	(-0.44)	(254.99)	(-0.15)	(-1.99)	(-0.65)
β_{20}	110.84	0.01	55.92	0.01	4.06	0.04
	(-105.28)	(-0.85)	(-74.78)	(-0.89)	(-20.15)	(-1.93)
β_{21}	30.27	0.08	572.83	0.09	0.22	0.07
	(-55.02)	(2.76)	(-239.34)	(2.99)	(4.68)	(2.69)
σ_1	232.54	0.00	57.17	0.00	28.79	0.01
	(152.49)	(-0.45)	(75.61)	(-0.31)	(53.66)	(-0.97)
σ_2	356.11	0.00	144.62	0.00	94.44	0.00
	(188.71)	(-0.12)	(120.26)	(-0.22)	(97.18)	(0.67)
			情景（c）			
β_{10}	62.573	0.001	334.622	0.026	0.007	0.116
	(-79.103)	(-0.24)	(-182.927)	(-1.607)	(0.841)	(-3.402)
β_{11}	669.793	0.046	1864.894	0.01	0.00	0.00
	(258.804	(-2.138)	(431.844)	(1.018)	(0.072)	(0.074)
β_{20}	15.413	0.003	35.746	0.001	0.713	0.034
	(39.259)	(0.584)	(-59.788)	(0.295)	(-8.445)	(-1.833)
β_{21}	1354.902	0.199	3457.934	0.05	0.002	0.094
	(-368.09)	(-4.46)	(-588.042)	(-2.227)	(-0.467)	(-3.062)
			情景（d）			
β_{10}	1.096	0.002	378.853	0.00	12.545	0.00
	(10.471)	(0.49)	(-194.642)	(0.213)	(35.419)	(-0.22)
β_{11}	46.717	0.00	2200.496	0.002	1.604	0.002
	(68.35)	(-0.139)	(469.094)	(-0.404)	(-12.666)	(-0.425)
β_{20}	219.251	0.00	43.403	0.004	29.719	0.012
	(-148.071)	(-0.188)	(-65.881)	(-0.654)	(-54.515)	(-1.1)
β_{21}	186.311	0.00	3031.079	0.001	0.006	0.001
	(-136.496)	(-0.112)	(-550.552)	(-0.292)	(-0.77)	(-0.277)
σ_1	1507.266	0.023	355.283	0.029	259.187	0.043
	(388.235)	(-1.526)	(188.49)	(-1.71)	(160.993)	(-2.068)
σ_2	642.017	0.137	342.25	0.064	95.851	0.05
	(253.381)	(-3.699)	(185.0)	(-2.535)	(97.904)	(-2.244)

表4： 拟合的混合比例的平均值（标准差）均方误差（数值乘以100）

n	GMLRs	CGMLR	GMoE	CG-MoE	S-G-MoE	S-CG-MoE
			情景（a）			
200	12.25 (0.35)	12.26 (0.35)	12.43 (0.47)	12.24 (0.26)	0.75 (0.48)	0.95 (0.63)
500	12.39 (0.14)	12.39 (0.14)	12.45 (0.18)	12.38 (0.10)	0.35 (0.21)	0.44 (0.26)
1000	12.76 (0.06)	12.76 (0.06)	12.80 (0.08)	12.76 (0.04)	0.20 (0.11)	0.25 (0.14)
			情景（b）			
200	24.4 (14.87)	13.15 (0.48)	28.88 (16.23)	13.07 (0.18)	0.91 (0.77)	1.14 (0.74)
500	29.5 (20.04)	12.27 (0.18)	30.17 (16.1)	12.16 (0.09)	0.6 (0.66)	0.55 (0.28)
1000	34.28 (21.62)	12.85 (0.08)	31.18 (13.19)	12.83 (0.04)	0.48 (0.61)	0.29 (0.17)
			情景（c）			
200	20.64 (14.74)	12.53 (0.72)	27.49 (13.41)	12.04 (0.17)	0.67 (0.54)	1.50 (1.03)
500	24.16 (18.28)	12.49 (2.69)	36.74 (12.59)	12.01 (0.08)	0.30 (0.19)	0.65 (0.40)
1000	29.79 (23.02)	12.24 (0.15)	38.23 (11.12)	12.18 (0.04)	0.27 (0.29)	0.37 (0.200)
			情景（d）			
200	39.03 (22.98)	13.09 (0.50)	35.14 (17.70)	12.94 (0.22)	1.66 (7.47)	1.10 (0.73)
500	43.46 (20.86)	12.02 (0.20)	35.67 (15.27)	11.86 (0.09)	0.72 (3.43)	0.49 (0.31)
1000	49.36 (23.80)	12.34 (0.14)	37.10 (15.94)	12.28 (0.06)	0.40 (0.53)	0.27 (0.16)

4 应用

在本节中，我们将在一个关于音调感知的真实数据集上展示所提出方法的实用性Cohen [1984]。研究人员向一名受训音乐家播放纯基音，并使用拉伸比添加电子生成的泛音。随后要求该音乐家将一个可调音调调至高于基音一个八度的位置。研究人员记录了同一名音乐家进行150次试验时的调音比和拉伸比。该数据集可在 `R` 软件包 `mixtools` Benaglia 等人 [2010]中找到。该数据集的图示见

1. 许多作者都对该数据集进行了混合分析。这些研究一致认为，该数据集包含两个成分。因此，在以下数据分析中，我们将使用 $K = 2$ 个成分。

首先，我们将所提出的 S-CG-MoLE 模型与以下模型进行比较：GMLRs (2)、污染 GMLRs (CGMLRs) Mazza 和 Punzo [2020], GMoLE Jacobs 等人 [1991], CG-MoLE (6)、半参数GMoLE (S- GMoLE) Huang and Yao [2012]。结果见表5。根据贝叶斯信息准则（BIC，-424.0539），数据支持 CGMLRs 模型为最佳模型。其次是

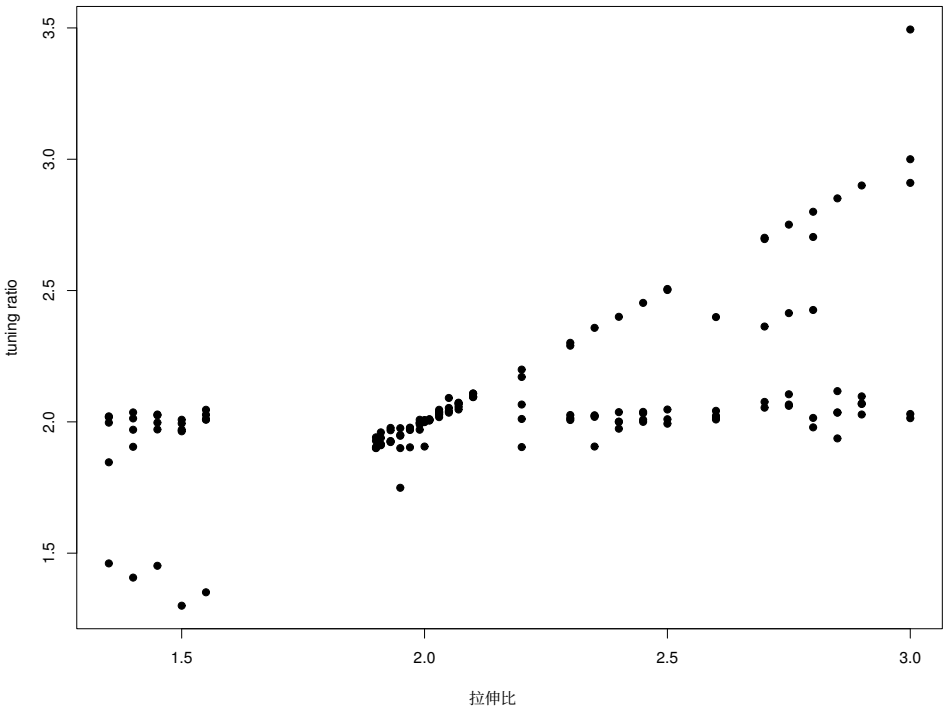


图1：音调感知数据的散点图

紧随其后的是 CG-MoLE (-419.0688) 和 S-CG-MoLE (-342.8331)。

图3绘制了从全部六个拟合模型中得到的拟合回归线。各组分的回归线几乎完全相同。在图2中，我们绘制了从所有拟合模型中得到的混合比例。不同拟合模型的结果各不相同。正如预期，参数化专家混合模型（GMoLE 和 CGMoLE）对混合比例给出了单调递增函数。值得注意的是，半参数回归混合模型（S-GMoLE 和 S-CG-MoLE）在拉伸比的不同取值处呈现出相反的模式。对于 S-GMoLE模型，当拉伸比介于1.5和2之间时，属于组分1的概率下降；当拉伸比介于2和2.5之间时，该概率又上升。S-CG-MoLE模型则呈现相反的关系。总体而言，根据贝叶斯信息准则，污染回归混合模型给出了最佳拟合。

为评估所提出方法在存在离群值时的性能，我们在音调感知数据中加入人工噪声。具体而言，我们将随机选取的 5% 数据点的 $y-$值乘以 2.5，从而污染数据。得到的受污染数据绘制于图4中。

表5：基于音调感知数据拟合的模型结果。

	GMLRs	CGMLR	GMoE	CG-MoE	S-G-MoE	S-CG-MoE
β_{10}	-0.0193	0.0034	-0.0295	0.0034	0.0060	0.0033
β_{20}	1.9164	1.9542	1.9132	1.9540	1.9250	1.9603
β_{11}	0.9923	0.9988	0.9957	0.9988	0.9832	0.9988
β_{21}	0.0426	0.0282	0.0437	0.0283	0.0393	0.0263
π_1	0.3023	0.4454	-	-	-	
σ_1	0.1328	0.0042	0.1373	0.0042	0.1211	0.0045
σ_2	0.0462	0.0244	0.0471	0.0244	0.0429	0.0247
γ_{10}	-	-	-0.7921	-0.7409	-	-
γ_{11}	-	-	2.6787	0.1559	-	-
η_1	-	3506.822	-	3509.089	-	2187.006
η_2	-	6.6516	-	6.6620	-	7.4175
α_1	-	0.7732	-	0.7737	-	0.7303
α_2	-	0.5553	-	0.5555	-	0.6678
BIC	-247.3224	$\mathbf{-424.0539}$	-245.6109	-419.0688	-105.8673	-342.8331

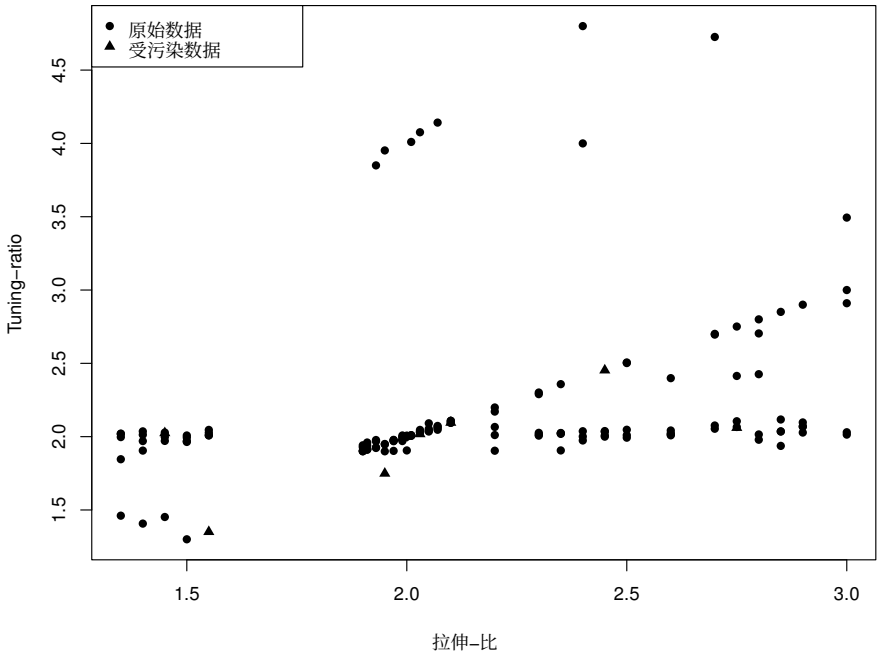


图4：受污染的声调感知数据的散点图

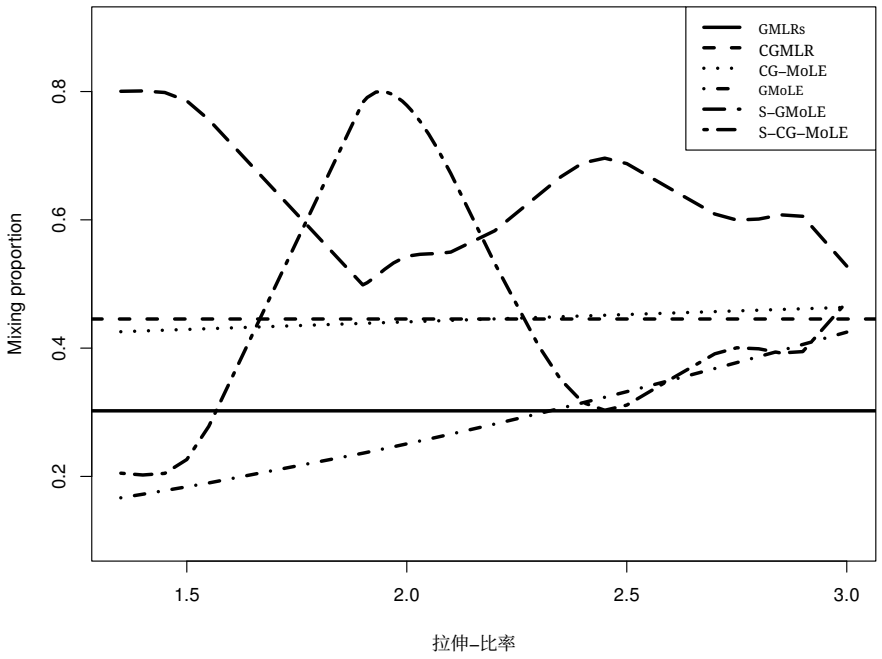


图2：音调感知数据中六个拟合模型第一成分的估计混合比例

我们将六个模型拟合到受污染的音调感知数据上。结果见表6。污染回归混合模型（CGMLR、CG-MoLE 和 S-CG-MoLE）的表现远远优于高斯回归混合模型（GMLRs、GMoLE 和 S-GMoLE）。图 6 展示了所有拟合模型的分量回归线，也印证了这一点。左图绘制了 GMLRs、GMoLE 和 S-GMoLE 模型，右图绘制了 CGMLR、CG-MoLE 和 S-CG-MoLE 模型。从图中可以看出，前一组模型会受到离群值的影响，而离群值似乎不会影响后一组模型。比较图3和图 6可见，拟合得到的污染回归混合模型并无差异。这凸显了污染模型在存在离群值时的稳健性。最后，我们在图5中绘出了所有拟合的混合比例。值得注意的是，所有污染模型的混合比例呈现出相同的关系，与数据中不存在离群值时得到的结果一致。

表6：在受污染的声调感知数据上拟合的模型结果

	GMLRs	CGMLR	GMoE	CG-MoE	S-G-MoE	S-CG-MoE
β_{10}	0.9885	0.0040	1.0664	0.0041	1.3100	0.0040
β_{20}	1.9090	1.9565	1.9071	1.9242	1.9368	1.9568
β_{11}	0.7396	0.9986	0.7061	0.9985	0.3515	0.9985
β_{21}	0.0457	0.0276	0.0465	0.0392	1.0327	0.0275
π_1	0.2579	0.4513	-	-	-	-
σ_1	0.8414	0.0042	0.8528	0.0042	0.0539	0.0042
σ_2	0.0506	0.0284	0.0510	0.0463	0.0417	0.0238
γ_{10}	-	-	-3.5595	114.3463	-	-
γ_{11}	-	-	1.1150	-38.0553	-	-
η_1	-	94535.212	-	69413.645	-	91350.613
η_2	-	7.7132	-	1162.205	-	7.5836
α_1	-	0.7020	-	0.8311	-	0.7008
α_2	-	0.5301	-	0.8825	-	0.5330
BIC	-78	$\mathbf{-292.9539}$	-79.9801	-278.5077	355.4877	-203.4893

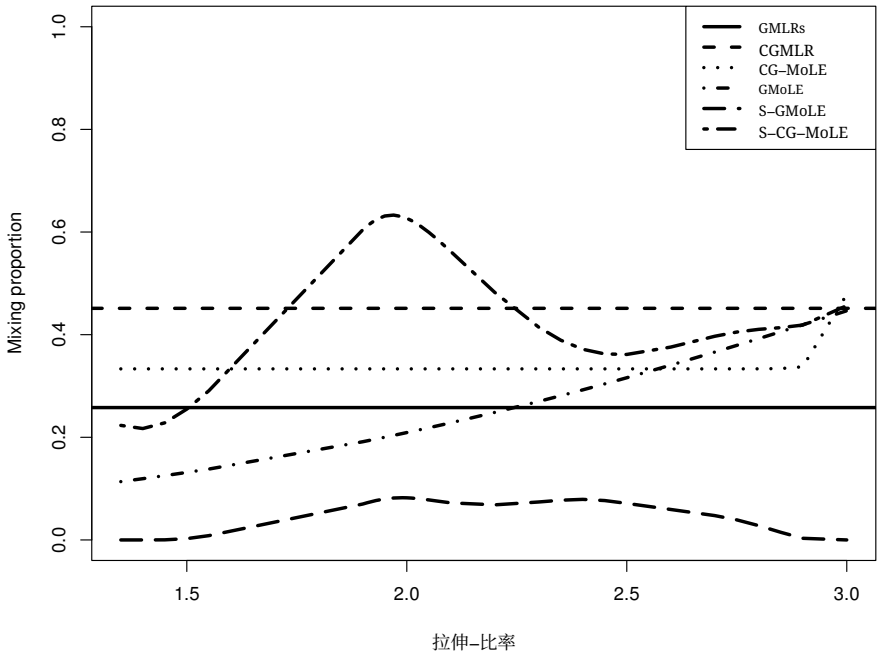
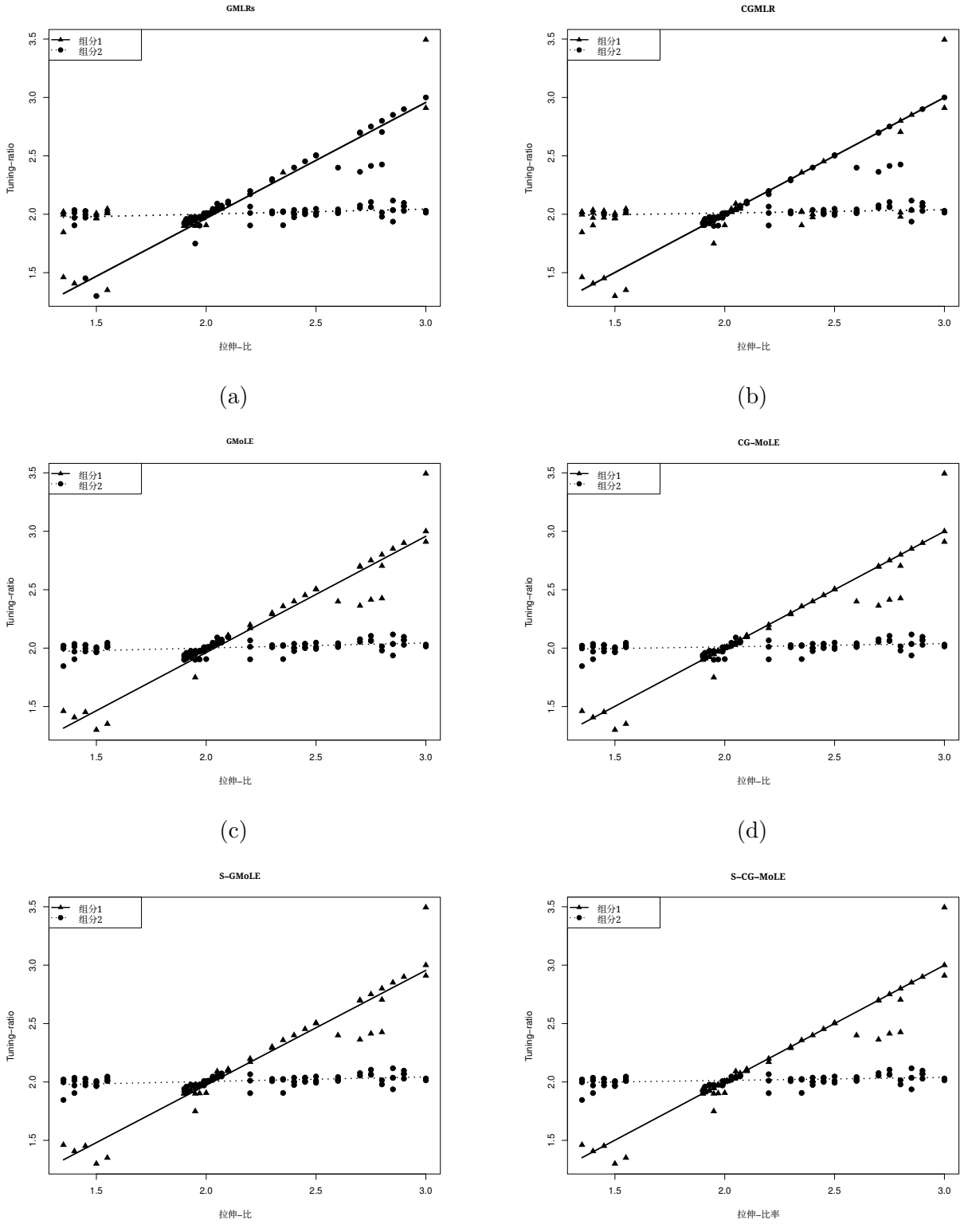


图5：六个拟合模型在受污染的声调感知数据上对第一成分的估计混合比例

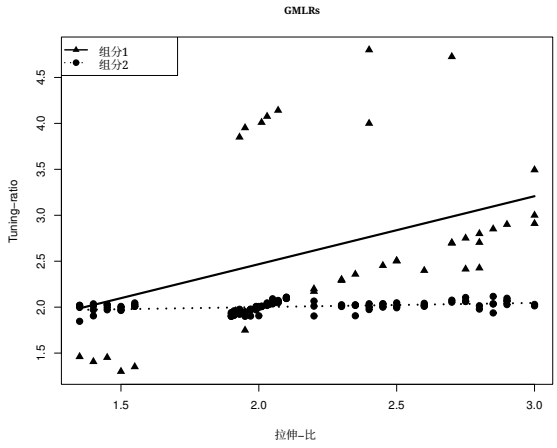


(a) (b)

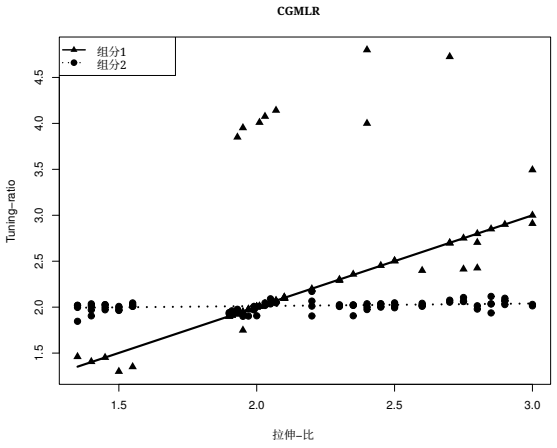
(c) (d)

(e) (f)

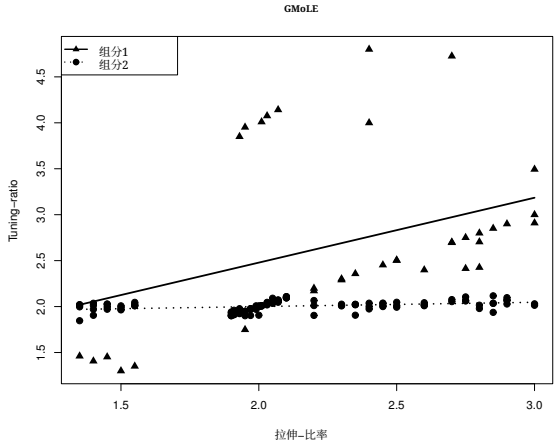
图3：音调感知数据上的估计模型：(a) GMLRs (b) CGMLR (c) GMoE (d) CG-MoE (e) S-G-MoE (f) S-CG-MoE



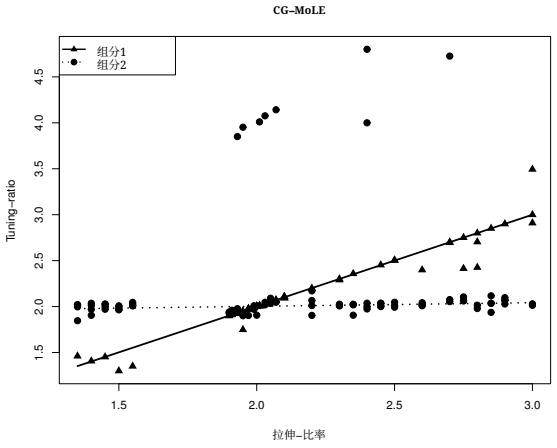
(a)



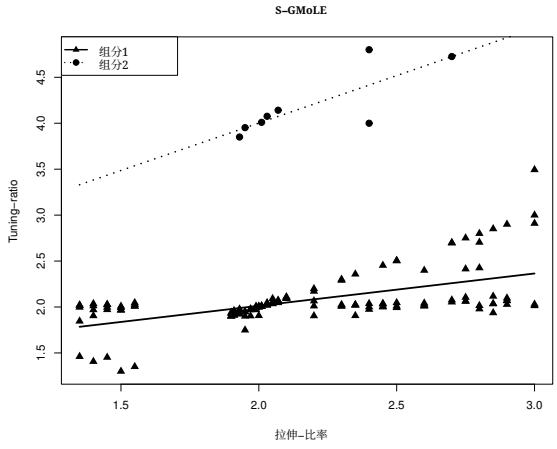
(b)



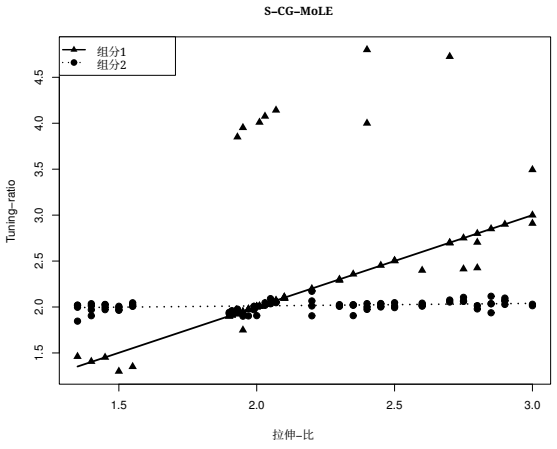
(c)



(d)



(e)



(f)

图 6：受污染的音调感知数据上的估计模型：(a) GMLRs (b) CGMLR (c) GMoE (d) CG-MoE (e) S-G-MoE (f) S-CG-MoE

5 结论

本文介绍了一种基于污染高斯专家（或成分）和灵活门控函数（或混合比例）的稳健专家混合模型。更具体地说，假设混合比例函数是协变量的光滑、未知且因此为非参数函数，并假设每个成分均服从污染高斯分布。本文提出通过期望条件最大化（ECM）算法进行最大似然估计，以估计该模型。非参数混合比例函数采用局部线性核估计量进行估计。通过广泛的模拟研究以及在真实数据集上的应用，证明了所提方法在存在轻度离群值时的实用性和稳健性。

为了避免局部核估计中维数灾难所带来的实际影响（即随着维数增加，局部区域中的数据变得稀疏），我们假设混合比例函数是单维协变量的函数。然而，通过采用加性模型或神经网络等机器学习方法，该模型可以轻松扩展以适应多维协变量Xue and Yao [2022]或随机森林。

声明

- 利益冲突/竞争性利益：作者声明不存在利益冲突。
- 数据可用性：支撑本研究结论的数据已公开可用。

参考文献

T. Benaglia、D. Chauveau、D. R. Hunter 和 D. S. Young。mixtools：用于分析混合模型的 R 软件包。统计软件杂志，32：1–29，2010年。C. M. Bishop。模式识别与机器学习。Springer，纽约州纽约市，2006年。F. Chamroukhi。使用 t 分布进行稳健专家混合建模。神经网络，79：20–36，2016年。E. A. Cohen。非谐泛音对音程感知的若干影响。音乐感知，1(3)：323–349，1984年。A. P. Dempster、N. M. Laird 和 D. B. Rubin。通过 EM 算法从不完整数据中进行最大似然估计。英国皇家统计学会杂志：*B*辑（方法论），39(1)：1–22，1977年。J. Fan。设计自适应非参数回归。美国统计学会杂志，87(420)：998–1004，1992年。J. Fan 和 I. Gijbels。局部多项式建模及其应用。CRC Press，纽约，1996年。Y. Ge、S. Xiang 和 W. Yao。基于 t 分布的稳健半参数混合回归模型。统计学通讯——模拟与计算，第 1–21 页，2024年。P. J. Green 和 B. W. Silverman。非参数回归与广义线性模型：粗糙度惩罚方法。CRC Press，1994年。M. Huang 和 W. Yao。混合比例变化的回归模型混合：一种半参数方法。美国统计学会杂志，2012年。doi：10.1080/01621459.2012.682541。M. Huang、R. Li 和 S. Wang。非参数回归模型混合。美国统计学会杂志，2013年。doi：10.1080/01621459.2013.772897。R. A. Jacobs、M. I. Jordan、S. J. Nowlan 和 G. E. Hinton。自适应局部专家混合模型。神经计算，3(1)：79–87，1991年。

A. Mazza 和 A. Punzo。多元污染正态回归模型混合。统计论文，61(2):787–822，2020年。X. Meng 和 D. B. Rubin。通过 ECM 算法进行最大似然估计：一个通用框架。*Biometrika*，80(2):267–278，1993 年。E. Mirfarah、M. Naderi 和 D. Chen。删失数据的线性专家混合模型：一种采用正态分布尺度混合的新方法。计算统计与数据分析，158:107182，2021年。T. Nagler 和 C. Czado。利用简化藤 copula 规避非参数密度估计中的维数灾难。多元分析杂志，151:69–89，2016年。H. D. Nguyen 和 G. J. McLachlan。拉普拉斯线性专家混合。计算统计与数据分析，93:177–191，2016年。R 核心团队。*R*：统计计算语言与环境。R统计计算基金会，维也纳，奥地利，2023年。网址 https://www.R-project.org/。S. B. Skhosana、S. M. Millard 和 F. H. J. Kanfer。用于估计非参数回归的半参数混合的修正 EM 型算法。统计与计算，34(4):125，2024年。Weixing Song、Weixin Yao 和 Yanru Xing。基于拉普拉斯分布的稳健混合回归模型拟合。计算统计与数据分析，71:128–137，2014年。R. Tibshirani 和 T. Hastie。局部似然估计。美国统计学会杂志，82(398):559–567，1987年。H. Wu 和 Jin-Ting Zhang。纵向数据分析的非参数回归方法：混合效应建模方法。约翰·威立父子公司，2006年。S. Xiang 和 W. Yao。非参数回归的半参数混合。统计数学研究所年刊，2018年。doi: 10.1007/s10463-016-0584-7。S. Xiang 和 W. Yao。带单指标的半参数回归混合模型，用于基于模型的聚类。数据分析与分类进展，2020年。doi: 10.1007/ s11634-020-00392-w。

J. Xue 和 W. Yao。嵌入机器学习的半参数回归混合模型及随协变量变化的混合比例。计量经济学与统计学，22:159–171，2022年。W. Yao、Y. Wei 和 C. Yu。使用 t分布的稳健混合回归。计算统计与数据分析，71:116–127，2014年。D. S. Young 和 D. R. Hunter。具有依赖于预测变量的混合比例的回归混合模型。计算统计与数据分析，54(10):2253–2266，2010年。C. B. Zeller、C. R. B. Cabral 和 V. H. Lachos。基于偏态正态分布的尺度混合的稳健混合回归建模。*Test*，25:375–396，2016年。