arXiv:2602.04788v1 [stat.ME] 4 Feb 2026

重新审视物种敏感性分布：一种贝叶斯非参数方法

Louise Alamichel[1*], Julyan Arbel[2], Guillaume Kon Kam King[3], IgorPrünster[1]

[1]博科尼数据科学与分析研究所，博科尼大学，*20136* 米兰，意大利[2]格勒诺布尔阿尔卑斯大学、法国国家科学研究中心、法国国家信息与自动化研究所、格勒诺布尔国立综合理工学院、*LJK*，*38000* 格勒诺布尔，法国[3]巴黎-萨克雷大学、法国国家农业、食品与环境研究院、*MaIAGE*，*78350* 茹伊昂若萨，法国

摘要

我们通过将物种敏感性分布（SSD）方法置于贝叶斯非参数（BNP）框架下，提出了一种生态风险评估新方法。SSD在全球范围内受到环境监管机构的广泛要求，但由于其在物种变异性建模中历来依赖参数假设，因此一直受到批评。我们采用非参数混合模型来克服这一局限，为SSD奠定了统计上稳健的基础。我们的BNP方法具有多项优势，包括能够有效处理小型数据集或删失数据（这类数据在生态风险评估中十分常见），并能在进行密度估计和聚类的同时，以有据可循的方式量化不确定性。我们选用一种特定的非参数先验作为混合测度，因其具有稳健的聚类特性；鉴于我们对组分数量缺乏强烈的先验认识，这一点至关重要。通过模拟研究和真实数据集分析，我们证明了BNP-SSD优于经典SSD方法。我们还提供了BNP-SSD Shiny应用程序，使生态毒理学界能够使用我们的方法。此外，我们利用混合模型固有的聚类结构，探索物种敏感性的模式。研究结果凸显了所提方法在改进生态风险评估方法方面的有效性。

关键词：贝叶斯非参数方法、临界效应浓度、生态风险评估、生态毒理学、危险浓度、混合模型。

*通讯作者：louise.alamichel@unibocconi.it

1 引言

生态毒理学背景。评估物种群落对环境胁迫的响应，对生态风险评估至关重要。为此开发的方法在复杂程度和现实性方面差异很大。物种敏感性分布（SSD）是一种中间层级的方法，比基础的评估因子（Posthuma 等人，2002年）更精细，但又足够实用，可供大多数发达国家的环境管理者和监管机构常规使用（澳大利亚和新西兰，ANZECC，2000年；加拿大，CCME，2007；中国，Liu 等人，2014；欧盟，欧洲化学品管理局，2008；南非，Roux 等人，1996；美国，美国环保署，1998）。物种敏感性分布方法旨在针对某种污染物，利用从一组物种中收集的信息，描述所有可能暴露于该污染物的物种的耐受性。这些信息包括临界效应浓度（CEC），即针对特定物种的浓度阈值；超过该阈值，物种就会受到严重影响。例如，这类效应水平可以是导致50% 受试生物死亡的浓度（半数致死浓度50% （LC_{50}）），也可以是相较于对照组抑制生长或繁殖50% 的浓度（半数效应浓度50% （EC_{50}））。每个临界效应浓度都是针对单一物种开展的高成本生物测定实验的汇总结果，因此数据通常十分有限。欧洲化学品管理局（ECHA）规定，最低数据要求为样本量10，最好达到15（欧洲化学品管理局，2008）。

现有参数方法的局限性。为了描述所有待保护物种的耐受性，需要根据受试物种样本估计临界效应浓度的分布。实践中，通常采用参数分布假设（Forbes 和 Calow，2002年）：假设临界效应浓度服从对数正态（Wagner 和 Lokke，1991；Aldenberg 等人，2002年）、对数逻辑斯蒂（Aldenberg 和 Slob，1993；Kooijman，1987）、三角（Van Straalen，2002；Zhao 和 Chen，2016）或 BurrIII（Shao，2000年）分布。确定群落响应特征后，便可定义安全浓度，通常采用物种中5% 的危险浓度（HC_5），即该分布的第5百分位数。实践者通常不会直接采用原始估计值，而是取置信区间的下限，并进一步用安全系数进行修正。由于难以证明选择某一种参数分布的合理性，许多研究转而探讨分布比较（Xu 等人，2015年；He 等人，2014年；Jagoe 和 Newman，1997年；Van Straalen，2002；Xing 等人，2014年；Zhao 和 Chen，2016）。这些研究形成的普遍共识是，没有哪一种分布在所有数据集中都能提供始终更优的拟合，最合适的模型在很大程度上取决于具体数据集（Forbes 和 Calow，2002年）。此外，由于样本量较小时拟合优度检验的统计功效往往较低，实际考量常常比理论因素更具决定性。在这种情况下，对数正态分布已成为惯常选择，主要是因为它在数学上简单、易于解释，并且便于估计HC_5的置信区间。

替代方案与对灵活性的需求。另一研究方向旨在完全避免参数假设，采用无分布方法，例如经验分布函数（Suter II 等，1999年；Jones 等，1999年）、基于秩的方法（Van Der Hoeven，2001年；Chen，2004年）、自助重抽样（Jagoe 和 Newman，1997年；Grist 等，2002年；Wang 等，2008年）或非参数核密度估计（Wang 等人，2015）。然而，这些方法需要较大的样本量，这与生态毒理学数据的稀缺以及减少动物试验的伦理目标相矛盾。这一局限凸显了模型普适性与数据可用性之间持续存在的权衡：尽管灵活的非参数方法

原则上具有吸引力，但其可靠性在数据稀疏或噪声较大时会急剧下降。另一条研究路线认为，临界效应浓度的分布可能并非同质，而是由若干亚组构成，每个亚组具有各自的分布形态（Zajdlik 等人，2009；Kefford 等人，2012；Craig，2013）。这在生态学上是符合实际的：分类学、栖息地或污染物的作用机制可能引起多峰性，尤其是针对特定分类群的化合物（例如某些农药）。然而，当怀疑存在多峰性时，现行官方建议是仅保留最敏感组（欧洲化学品管理局，2008；Zajdlik 等人，2009），这是一种务实但浪费信息的策略，过度简化了生态变异性。这些考量进一步表明，需要能够灵活适应复杂、多峰且数据匮乏情形的方法，以便在不施加限制性分布假设的情况下刻画异质性敏感性模式。

采用贝叶斯非参数方法的理由。 鉴于物种敏感性决定因素存在不确定性，关于群组结构的先验知识很少，因此需要采用非参数方法。在此类生态学环境中，物种及其敏感性之间的关系十分复杂，通常无法通过固定的参数形式加以刻画。贝叶斯非参数（BNP）模型尤其具有吸引力，因为它们能够让数据揭示潜在结构，而无需对敏感性组的数量或形态施加限制性假设。现有的频率学派非参数方法依赖渐近保证，因此在小型数据集上的表现可能较差。贝叶斯非参数混合模型能够有效适用于大样本和小样本，因为混合模型的复杂度会根据可用数据进行调整。贝叶斯非参数模型允许根据数据学习混合成分的数量，而不是预先固定，从而避免在生态学应用中难以论证的任意假设。它们还能够自然地量化关键监管量（例如 $HC_{5,}$ ）的不确定性，提供完整的后验分布，而非点估计。与核密度方法相比，它们能够降低过拟合风险，并提供合理的不确定性量化（Barrios 等人，2013年）。重要的是，小型数据集中的有限信息可以通过信息先验得到补充，借助来自其他物种或污染物的外部知识（Awkerman 等人，2008年；Craig，2013年；Craig 等人，2012年）。通过这种方式，贝叶斯非参数混合模型将生态学现实性（通过灵活的多峰结构）与监管需求（通过可靠的不确定性量化）相结合，因此特别适用于物种敏感性分布应用。

目标与提纲。 在第 2 节，我们提出一种基于非参数混合模型的物种敏感性分布贝叶斯非参数方法。我们表明，我们的贝叶斯非参数物种敏感性分布方法可以纳入删失数据，这类数据在生态毒理学中很常见（Kon Kam King 等人，2014），并且能够对感兴趣变量 HC_5 的不确定性给出严格描述。我们通过将 BNP-SSD 与最标准的正态物种敏感性分布方法（Aldenberg 和 Jaworska，2000）以及基于核密度估计（KDE）的非参数物种敏感性分布方法（Wang 等人，2015）进行比较，展示本方法的价值。该比较在第 3 节的模拟数据上进行，以证明 BNP-SSD 具有更高的准确性；我们还在第 4 节研究真实的删失与非删失数据集，从而凸显本方法的稳健性。此外，我们进行探索性分析[†]，以描述借助 BNP-SSD、通过研究第 5 节中由聚类所诱导的物种或污染物模式能够获得哪些生物学见解。最后，我们在第 6 节作出结论并阐述进一步的研究方向。

[†]代码可在 GitHub 上获取，地址为 https://github.com/alamichL/BNP_物种敏感性分布/。

2 方法

由于变异幅度很大的物种敏感性分布浓度均为正值，通常的做法是对浓度进行对数变换。我们考虑一个容量为 n 的对数浓度样本，记为 $\boldsymbol{X} = (X_1, \ldots\ldots, X_n)$，它通常表示用给定污染物测试的、由 n 个物种组成的集合的 CEC。此外，数据经过标准化：观测值被中心化并重新缩放，使其方差为 1。推断之后，所有估计均变换回原始尺度。

我们基于样本进行SSD的密度估计 $\boldsymbol{X}$，采用贝叶斯非参数混合模型。以非参数先验作为混合测度的概率密度核混合方法源自Lo（1984），其介绍了狄利克雷过程混合模型（DPM）。DPM的推广形式允许混合分布采用任意离散非参数先验。通过对称为完全随机测度（Kingman，1967）的随机测度进行归一化，可以得到一大类此类先验分布。归一化步骤产生了所谓的独立增量归一化测度（NRMI），其定义见Regazzini 等人（2003）。综述请参见Lijoi 和 Prünster（2010）；Jordan（2010）；Barrios 等人（2013）。后文将进一步介绍NRMI先验的具体选择及其推断影响。NRMI混合模型定义为

$$\begin{aligned} X_i \mid \tilde{P} &\overset{\text{i.i.d.}}{\sim} \tilde{f}(x) = \int k(x \mid \theta)\tilde{P}(\mathrm{d}\theta), \\ \tilde{P} &\sim \text{NRMI} \end{aligned} \tag{1}$$

其中，k 是由某些 $\theta \in \Theta$ 参数化的概率密度核，而 P 是定义在 Θ 上的随机概率，其分布为独立增量归一化测度（NRMI）。或者，也可以将混合模型表示为层级形式：

$$\begin{aligned} X_i \mid \theta_i &\overset{\text{ind}}{\sim} k(x \mid \theta_i), \quad i = 1, \ldots, n, \\ \theta_i \mid \tilde{P} &\overset{\text{i.i.d.}}{\sim} \tilde{P}, \quad i = 1, \ldots, n, \\ \tilde{P} &\sim \text{NRMI}. \end{aligned} \tag{2}$$

具体而言，我们考虑位置-尺度混合，并将 $\theta_i = (\mu_i, \sigma_i)$ 记为个体均值和标准差向量，$\theta_i \in \mathbb{R} \times \mathbb{R}_+$。正如引言中所述，对数浓度通常采用正态分布或此类分布的混合进行拟合。我们的目标是从这些参数模型转向（2）中的非参数设定；为了能够与其他方法进行比较，我们对对数浓度 k 采用正态设定，$\boldsymbol{X}$，$k(x \mid \mu, \sigma) = \mathcal{N}(x \mid \mu, \sigma^2)$。

在此框架之 中，密度估计通过计算后验均 n

$$\hat{f}(x \mid \boldsymbol{X}) = \mathbb{E}\big(\tilde{f}(x) \mid X_1, \ldots, X_n\big) \tag{3}$$

对于任意 x 在 $\mathbb{R}$。对于后验抽样，我们使用`BNPdensity`[‡]，这是一个 R 软件包，可基于广义伽马过程的一般 NRMI 先验设定进行贝叶斯非参数密度估计和聚类（参见Barrios 等人，2013年；Arbel 等人，2021）。`BNPdensity`采用了广受欢迎的 Ferguson–Klass 算法（Ferguson 和 Klass，1972），并将其扩展为 Metropolis–Hastings-within-Gibbs 方案。

[‡]可于https://CRAN.R-project.org/package=BNPdensity获取。

我们将所提出的 BNP-SSD 与两种竞争方法进行比较。第一种是 SSD 最常用的模型——正态分布（Aldenberg 和 Jaworska，2000年），其估计密度为 $\hat{} \ f_{\mathcal{N}}(x) = \mathcal{N}(x \mid \hat{}\mu, \hat{}\sigma^2)$，其中 $\hat{}\mu$ 和 $\hat{}\sigma$ 分别是数据的经验均值和标准差。第二种是 Wang 等（2015年）近期应用于 SSD 的频率学派非参数核密度方法，其估计密度为 $\hat{} \ f_{\mathrm{KDE}}(x) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{N}(x \mid X_i, h_n^2)$，其中 $h_n = 1.06\hat{}\sigma n^{-\frac{1}{5}}$ 是 Silverman（1986）提出的渐近最优默认带宽建议，Wang 等（2015年）也采用了这一建议。

2.1 删失数据

Kon Kam King 等人（2014）阐述了如何在正态SSD中使用删失数据，并展示了SSD中常见做法的缺陷：对删失数据进行变换或丢弃。类似地，也可以将删失数据纳入BNP-SSD。

事实上，只需对（2）中定义的层次模型的第一行作适当调整，而层级结构中的其他层级保持不变。具体而言，设 F 为核 k的累积分布函数，则： k（$x \mid \theta$）需替换为F（$x \mid \theta$），用于左删失观测；替换为 $1 - F$（$x \mid \theta$），用于右删失观测；对于区间删失观测F（$x_r \mid \theta$）$- F$（$x_l \mid \theta$），则替换为 $[x_l, x_r]$。此方法可用于标准正态SSD和任何基于似然的推断，但目前没有广泛可用的工具能够对所有类型的删失数据进行核密度估计：R包`ICE` 无法处理左删失/右删失数据（已不再由CRAN维护）；R包`muhaz` 或`Kernelheaping`分别可处理右删失数据或区间删失数据，但似乎尚无可处理全部三种删失数据的实现。

本文以一个删失数据集为例，展示各种方法之间的差异；为便于比较，我们研究了该数据集的删失版本和未删失版本。要从删失数据集得到未删失数据集，我们采用常见做法：舍弃左删失和右删失数据，并用区间的中间值替代区间删失数据。

2.2 先验设定

NRMI 先验的类别非常广泛，详尽综述请参阅 Lijoi 和 Prünster (2010)。本文聚焦于该类别中的一个特定成员，即归一化稳定过程（Kingman, 1975）；正如 Barrios 等人 (2013所指出的，它是一个自然的默认选择。该过程由稳定性参数 $\gamma \in (0,1)$ 和基测度 P_0设定，后者对应于随机概率测度的期望。稳定性参数γ 控制簇数量先验分布的变异性：直观而言，较小的 γ 对应信息性先验，而较大的 γ 则表示先验较为宽泛。较大的 γ 值可能需要大量计算，才能保持后验抽样的质量。我们选择 $\gamma = 0.4$，以兼顾模型的灵活性和计算需求。详情请参阅 Lijoi 等人 (2007); Barrios 等人 (2013)。关于模型对 γ 选择的敏感性分析见图22。

如前所述，我们将考虑位置-尺度混合，这意味着 NRMI 先验应定义在 $\mathbb{R} \times \mathbb{R}_+$（即位置和尺度构成的空间）上。因此，基测度 P_0定义在 $\mathbb{R} \times \mathbb{R}_+$上，我们将其相对于 $\mathbb{R} \times \mathbb{R}_+$上的勒贝格测度的密度记为 f_0。我们假设位置 μ和尺度 σ先验独立。因此，我们使用

记号 $f_0(\mu,\ \sigma) = f_0^1(\mu)f_0^2(\sigma)$，其中可以为 f_0^1和 f_0^2指定超参数。能够为 μ 和 σ 指定独立先验，是 归一化随机测度的一项优点，因为先验无需具备任何共轭结构。因此，先验设定可以直接得出。

簇 μ 的位置参数所采用的具体分布为正态分布 $f_0^1(\mu) = \mathcal{N}(\mu \mid \varphi_1, \varphi_2^{-1})$，其中均值 φ_1 和精度 φ_2服从正态-伽马共轭超先验。这相当于对簇的位置采用了一个宽松的先验，因此簇可以位于数据集的中心，也可以位于边界。对于簇的尺度参数 σ ，鉴于预处理过程中对数据进行了标准化，使方差设为 1， σ 应小于 1，只有在单峰的情况下才接近 1。此外， σ 的先验下界也应受到限制，因为对于物种敏感性分布而言，大量极小的簇在生物学上并无意义。因此，我们选择均匀先验 $f_0^2(\sigma) = \text{Unif}_{[0.1,1.5]}(\sigma)$，并在上界 1 附近留出空间，以允许分布呈现潜在的单峰性。我们通过改变该先验设定的极值点，研究了结果对这一先验的敏感性；同时，也考察了均值为 0.5、方差为 1、下界为 0.1 的左截断正态先验。请注意，后者约有 $3/4$ 的概率质量位于 $\text{Unif}_{[0.1,1.5]}$ 分布的支撑集内。敏感性分析（见图 23）表明，对这两种先验分布的参数作适度调整，结果变化不大。

2.3 HC_5的估计

生态风险评估中主要关注的量是HC_5, ，它对应于物种敏感性分布的第5百分位数。在我们的贝叶斯非参数框架中，我们依赖于后验期望。在HC_5 上诱导的后验分布通过考虑（随机）累积分布函数 $\tilde{F}(x) = \int F_k(x|\theta)\, \tilde{P}(d\theta)$得到，其中 F_k 表示核 k的累积分布函数。累积分布函数的逆是分位数函数，因此为$\text{HC}_5 = \tilde{F}^{-1}(0.05)$。该反演在每次迭代时进行，即~针对马尔可夫链蒙特卡洛链中 F的每一次实现，从而使我们能够获得HC_5的后验分布。95%可信带由HC_5后验分布的2.5%和97.5%分位数构成。在实践中，我们的方法基于混合~~测度 P的截断近似，对累积分布函数的各次实现进行数值反演。对 P 的截断会得到有限混合。因此，其累积分布可以方便地求值，并且适合数值反演。截断方法是一种动态近似，旨在保持无限维混合测度的矩（Arbel 和 Prünster，2017）。

同样，KDE的第5百分位数可通过对累积分布函数进行数值反演得到，其置信区间则可通过非参数自助法获得。正态SSD的第5百分位数及其置信区间是按照经典方法得出的，该方法由Aldenberg 和 Jaworska（2000年）提出。

2.4 使用留一交叉验证进行稳健性比较

我们使用预测性能度量——留一（LOO）交叉验证，比较三种物种敏感性分布（SSD）方法的稳健性。我们按如下方式计算每种方法的留一交叉验证：

$$\text{LOO}_i = \hat{f}(X_i \mid \boldsymbol{X}_{-i}), \tag{4}$$

ˆ其中， $f(\cdot \mid \boldsymbol{X}_{-i})$ 是基于 $\boldsymbol{X}$ 计算的密度估计，其中对于三种方法，分别将 X_i 留出。留一交叉验证通过评估排除某个观测值后方法对该观测值的预测效果，来衡量方法的稳健性。过拟合的方法往往会因特定数据点的纳入或排除而产生显著变化的预测，导致留一交叉验证表现不佳。相比之下，稳健的方法无论是否移除单个数据点，都应能给出稳定的预测，从而取得更好的留一交叉验证表现。

对于贝叶斯非参数模型，LOO 被称为条件预测序数（CPO）统计量。它们在应用中十分常用，例如参见Gelfand（1996）。

每个数据点 X_i 的 CPO 统计量定义为

$$\mathrm{CPO}_i = \hat{f}(X_i \mid \boldsymbol{X}_{-i}) = \int k(X_i \mid \theta)\pi(\mathrm{d}\theta \mid \boldsymbol{X}_{-i}),$$

ˆ其中 $\pi(\mathrm{d}\theta \mid \boldsymbol{X}_{-i})$ 是与 $\boldsymbol{X}_{-i}$ 相关联的后验分布，且 $f(\cdot \mid \boldsymbol{X}_{-i})$ 是（交叉验证的）后验预测分布，对应于（3）。CPO 可以很容易地通过蒙特卡罗方法近似为

$$\widehat{\mathrm{CPO}_i} = \left(\frac{1}{T} \sum_{t=1}^{T} \frac{1}{k(X_i \mid \theta^{(t)})} \right)^{-1},$$

其中 $\{\theta^{(t)}, t = 1, 2, \ldots, T\}$ 是通过 BNPdensity R 软件包获得的马尔可夫链蒙特卡罗样本。对于这两个频率学派模型，可通过将其直接拟合到每个 i的留一数据 $\boldsymbol{X}_{-i}$ 上来计算 LOO。

2.5 聚类估计

如何基于马尔可夫链蒙特卡洛后验样本估计数据聚类，是贝叶斯统计中一个由来已久的问题（参见 Dahl，2006；Lau 和 Green，2007）。由于划分空间的基数会随样本量 n极为迅速地增长，估计聚类结构的计算开销很大；这一基数称为阶数为 n的贝尔数。实际中无法枚举所有划分，因此通常只能采用近似方法。文献中已有许多临时方法。然而，正如 Dahl（2006）所指出的，在一个本身就会产生聚类的模型之上再应用临时聚类方法，似乎有悖直觉。

相反，我们采用完全贝叶斯方法，从决策论依据出发进行聚类。我们考虑一个损失函数 $\mathcal{L}$，并提出一个贝叶斯点估计量 ˆc，将其定义为给定数据 $\boldsymbol{X}$ 时使后验期望损失最小化的聚类结构。

$$\hat{c} \in \arg\min_{c'} \sum_{c} \mathcal{L}(c', c)\pi(c \mid \boldsymbol{X}), \tag{5}$$

其中，对 c 的求和涵盖所有可能的聚类，而 $\pi(c \mid \boldsymbol{X})$ 表示聚类 c 的后验分布。

文献中常考察的最大后验概率（MAP）是此类贝叶斯估计量的一个例子，它基于非常粗糙的 $0-1$ 损失函数 $\mathcal{L}_{0-1}$。然而，当 n 较大时，后验样本通常几乎不会两次访问同一个聚类，因此，马尔可夫链蒙特卡洛输出的经验最大后验概率对链的初始化非常敏感，在实践中的有效性也十分有限。

显然，还可以考虑许多其他损失函数，而且预计它们的表现会优于 $\mathcal{L}_{0-1}$。其中有一种损失函数尤为突出，在估计聚类中的组数方面表现最佳。这种损失函数称为信息变差，记作$\mathcal{VI}$，它是信息论中确立已久的一种损失函数（Meilă，2007；Wade 和 Ghahramani，2018）。两个聚类之间的信息变差，定义为它们各自的信息量（即香农熵）之和，减去二者共享信息量的两倍。模拟结果表明，信息变差是一个合理的选择：当 Binder损失（Binder，1978）等其他损失通常倾向于高估簇的数量时，信息变差似乎能够有效地恢复其数量（例如，参见Wade 和 Ghahramani，2018中的模拟示例，尤其是图 6–8）。

所提出方法的一个优点见于Wade 和 Ghahramani（2018），即它依赖于贪心搜索算法来确定（5）的最小损失聚类。从马尔可夫链蒙特卡洛输出出发，这种贪心方法探索划分空间，而不局限于马尔可夫链蒙特卡洛链所访问过的划分，从而找到最优解。已提出多种高效实现，例如Rastelli 和 Friel（2018）以及Dahl 等（2022）。在本研究中，我们选择使用由Dahl 等（2022）提出的算法。结果见于Dahl 等（2022），表明与其他实现相比，该算法具有更优的性能和更高的计算效率。

3 模拟研究

为比较 BNP-SSD、正态SSD（Aldenberg 和 Jaworska，2000年）以及非参数 KDE-SSD（Wang 等人，2015）的性能，我们使用对应于不同情景的合成数据集开展了一项模拟研究。

3.1 模拟数据

我们考虑三种不同的模拟数据情景，分别对应于各种情况：

(a) 标准正态分布：这是为物种敏感性分布所做的正态假设成立的情形；(b) 具有三个自由度且非中心参数等于 -2的 t分布：这是某些物种相对更为敏感、从而在分布左侧形成更重尾部的情形；(c) 对应于正态混合 $\frac{1}{3}\mathcal{N}(-2,1) + \frac{2}{3}\mathcal{N}(5,1)$的双峰分布：这是一组物种比所有其他物种敏感得多的情形，典型见于某些对目标物种产生不成比例影响的农药。

这三种情景代表了真实数据中经验分布的多样性，例如国家公共卫生与环境研究所（RIVM）数据库（de Zwart，2001年）。

在所有设定中，我们独立抽取 $S = 40$ 大小为 10、20、50 和 100 的数据集。这些数据集规模具有代表性，正如引言中所述。图1展示了数据生成密度，以及使用三种不同方法对大小为 20 的数据集得到的不同估计值。

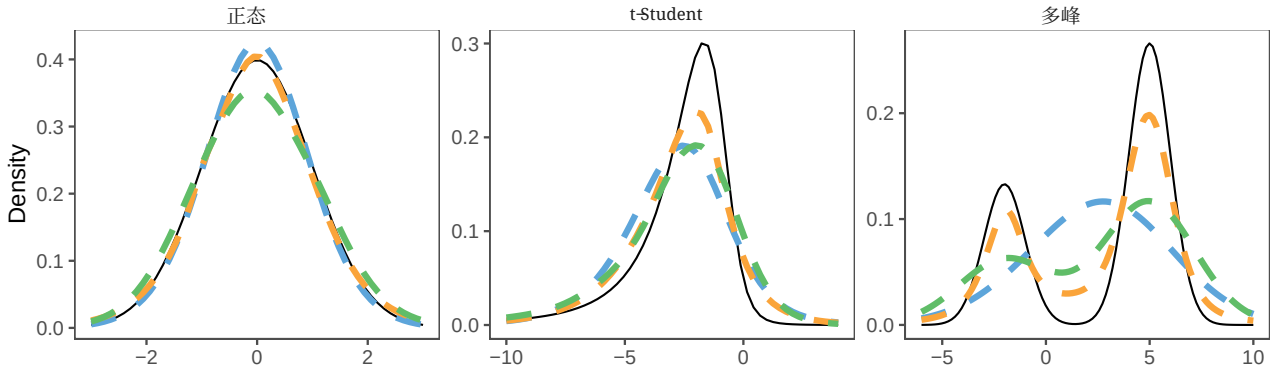

图 1： 三个模拟场景：数据生成密度（实线）以及基于样本量为 $n = 20$ 的数据集、由各模型得到的密度估计（虚线）。橙色（ ）表示贝叶斯非参数模型，蓝色（ ）表示正态模型，绿色（ ）表示 KDE 模型。

3.2 三种方法的性能比较

对于每个抽样数据集 $i \in \{1, \ldots, S\}$ 和所考虑的每个模型，我们估计数据生成密度 f，记为密度 f_i，并计算相应的第5百分位数 q_i，将其用作真实HC_5 的估计值，记为 q_0。我们将(l_i, u_i) 记为95% 置信区间/可信区间，适用于 q_i；并将 $\ell_i = u_i - l_i$ 记为其长度。为考虑抽样变异，我们计算平均汇总统计量（使用记号 $\langle \cdot \rangle_S$ 表示对这些 S 独立样本取平均）。

我们计算两个性能指标，即平均绝对误差 $= \langle |q_i - q_0| \rangle_S$，以及平均积分平方误差 $= \langle \int (f_i - f)^2 \rangle_S$。此外，我们计算平均置信区间/可信区间长度 $= \langle \ell_i \rangle_S$，作为不确定性的度量：由于BNP模型能够刻画模型不确定性，我们预期它会比其他模型给出更为保守的不确定性估计。然而，我们不希望保守到使这些估计对于估计HC_5这一实际目的毫无用处。

密度估计首先直观地表明，BNP-SSD在恢复真实密度方面优于另外两个模型（图1）。模拟研究的结果见图2，我们将按从上到下、从左到右的顺序加以说明。

对于正态模拟数据，设定正确的正态模型显然表现最佳。然而，贝叶斯非参数模型在 HC_5 上的平均绝对误差与正态模型非常相近。样本量较小时，贝叶斯非参数模型的平均积分平方误差几乎与正态模型相同。这说明，贝叶斯非参数模型的复杂度会随数据量变化；在数据匮乏情境下，它实质上会退化为正态模型。数据集规模较小时，贝叶斯非参数模型的平均置信区间长度大于正态模型，这反映了贝叶斯非参数模型中固有的模型不确定性，也表明随着更多数据的获得，该模型有望灵活适应偏离情况。数据集规模较大时，模型不确定性会降低，贝叶斯非参数模型与正态模型趋于一致。

对于 t-Student 模拟数据，BNP模型优于另外两种模型。正态模型和贝叶斯非参数模型的平均绝对误差小于核密度估计；贝叶斯非参数模型和核密度估计的平均积分平方误差较小，而正态模型的均值置信区间长度小得具有误导性。

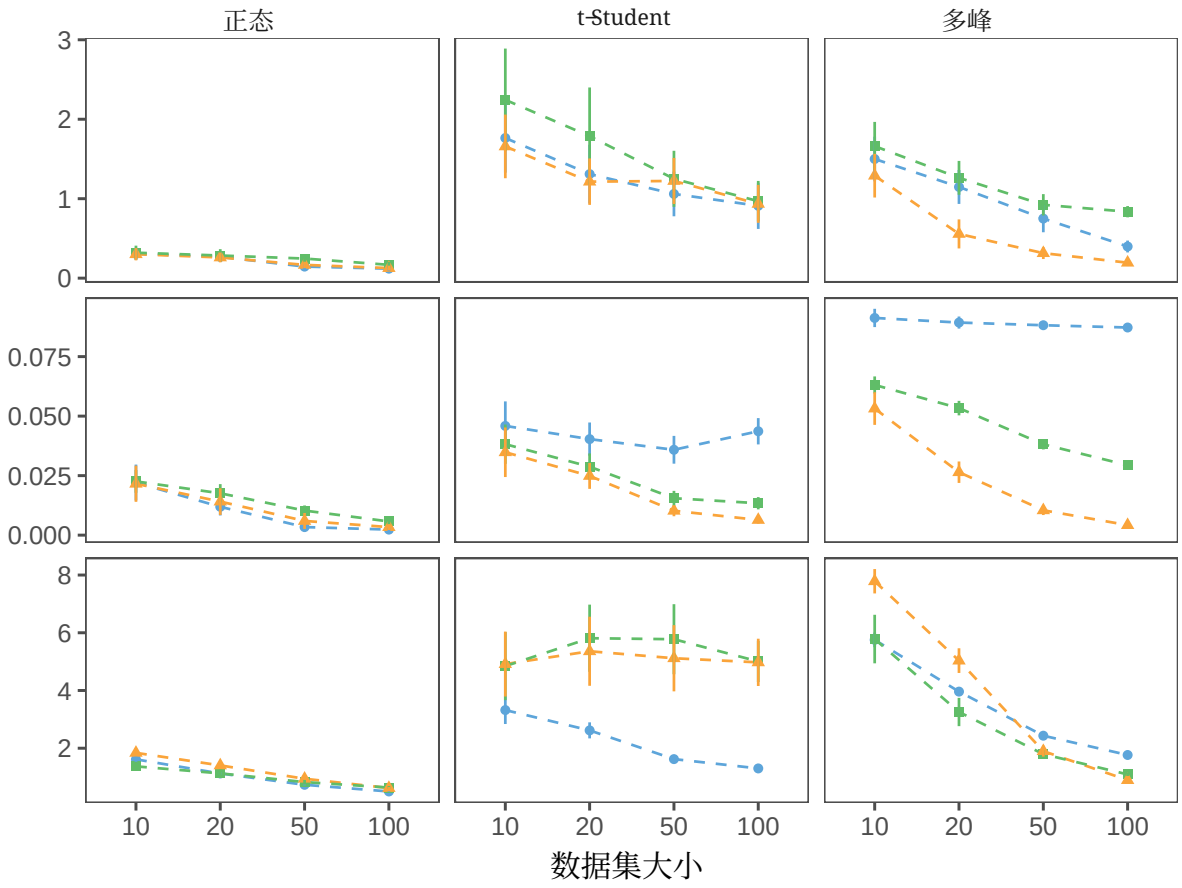

图2： 正态、t-Student分布和正态混合模拟场景（从左到右）；平均绝对误差（平均绝对误差）、平均积分平方误差（平均积分平方误差）和平均置信区间/可信区间长度（平均置信区间/可信区间长度）随数据集大小的变化（从上到下）。基于 $S = 40$ 次模拟估计的不确定性通过误差条表示。橙色（ ）表示贝叶斯非参数模型，蓝色（ ）表示正态模型，绿色（ ）表示核密度估计模型。

最后，对于多峰模拟数据，贝叶斯非参数模型在平均绝对误差和平均积分平方误差方面明显优于另外两种模型。数据集规模较小时，其平均置信区间长度相对另外两种模型较大；数据集规模较大时则较小。

4 污染物聚类分析

4.1 真实数据描述

我们通过从 RIVM 收集的大型数据库中选取若干污染物来说明所提出的贝叶斯非参数方法的优势，该数据库最初由deZwart（2001年）首次介绍。

我们研究由 Hickey等人（2012）整理的数据集，并沿用相同的数据质量和同质性限制。我们同时考虑该数据集的删失版本和非删失版本；非删失版本按照传统方法构建，即舍弃左删失数据和右删失数据，并取区间删失

数据的中心值。该数据集是一个水生生态毒理学研究数据库，包含1,557个物种和3,448种不同的化学品。具体而言，该数据集记录了以下协变量：物种、化学品和浓度。

为处理单个物种存在多个临界效应浓度值的情况，对于非删失数据，我们采用经典方法，以这些值的几何平均值作为替代值（欧洲化学品管理局，2008）；对于删失数据，则遵循 Kon Kam King 等人（2014）的方法。

4.2 密度、分位数和HC_5 估计

为便于说明，我们展示了三类污染物：大型数据集（包含 50 个以上的值）、中型数据集（约有 25 个值），以及小型数据集（略多于 10 个值）。对于每一类，我们都选取与某种污染物相关的数据集，其分布表现出近似单峰、偏斜和双峰特征，与模拟研究中的情况相同。这些数据集均为未删失数据集。我们对每个数据集拟合了三种模型（贝叶斯非参数、核密度估计和正态模型），并研究了HC_5的估计值及其可信区间、留一法误差，以及估计密度的形状与直方图的比较。我们还使用贝叶斯非参数模型和正态模型研究了同一批数据集的删失版本；而对于删失数据，似乎尚无核密度估计模型的实现（见第2.1节）。结果见附录A，参见图7至图12。

BNP模型比KDE 模型更加灵活，这一点从双峰数据集的密度估计中可以看出；同时，它的稳健性相当，甚至更强。三种方法所得置信区间/可信区间的长度并无显著差异。这有力地证明，限制较少（就分布假设而言）的模型并不会导致 HC_5估计过于保守。这一点至关重要，因为过于保守的估计会严重妨碍贝叶斯非参数物种敏感性分布方法的广泛应用。更具体地说，对于近似正态的数据集，贝叶斯非参数方法所得 HC_5 估计与正态模型的估计相当。当数据集显著偏离正态模型时，正态模型与贝叶斯非参数模型所得 HC_5估计差异显著，充分说明相较于正态模型，更灵活的贝叶斯非参数模型更值得采用。

4.3 按污染物聚类

迄今为止，我们已经展示了贝叶斯非参数方法在密度估计和确定HC_5方面相较于现有物种敏感性分布方法的优势。贝叶斯非参数物种敏感性分布方法还有一项额外优势：混合模型会对物种进行聚类，从而提供从生物学角度看颇有意义的信息。事实上，关于物种敏感性分布以及整个生态毒理学领域，一个长期存在的问题是，究竟是什么决定了物种对某种污染物的敏感性。Craig（2013年）认为分类学是一个决定因素，并据此有效地按照分类单元进行聚类。de Zwart（2001年）通过比较淡水物种和咸水物种，研究栖息地的影响；而Kefford 等人（2012年）则研究世界不同地区物种敏感性的差异。

所有这些方法都从一种可能的聚类结构出发，并检验各聚类单元之间是否存在显著差异。BNP-SSD 则采取相反的路径，将聚类

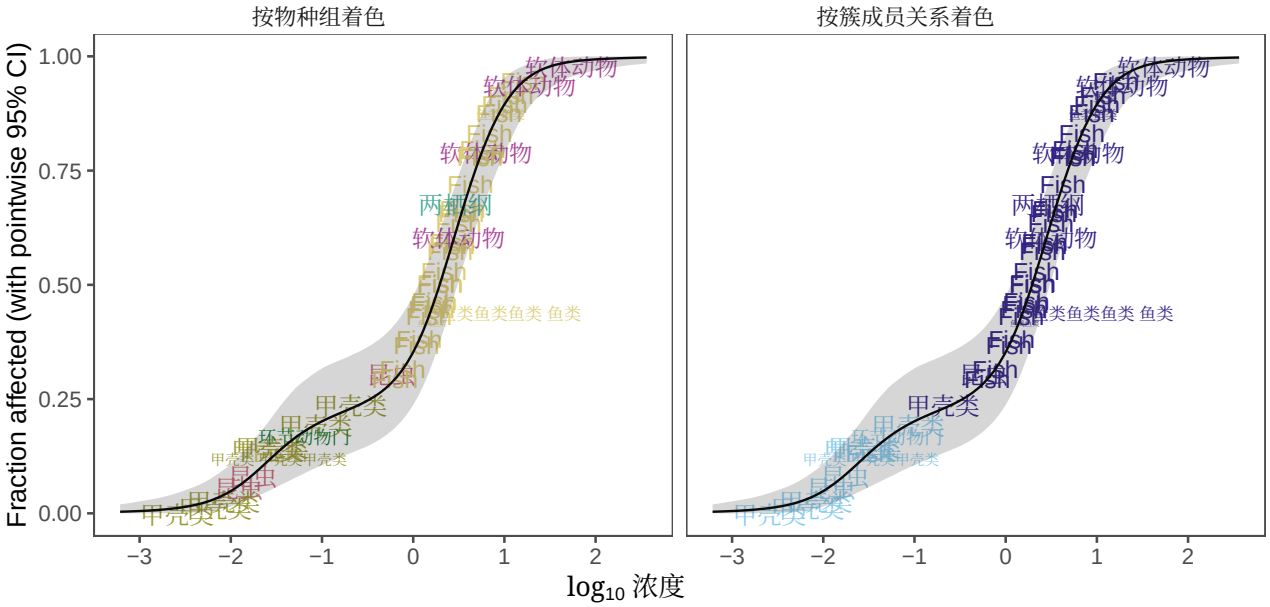

图 3：甲萘威（CAS：63-25-2）的物种敏感性分布，叠加显示各物种所属的准分类群以及累积分布函数的估计值（实线）。左：按准分类群为物种着色。右：按 BNP 模型中的簇成员关系为物种着色。浅灰色表示95%点态可信带，根据 BNP 模型的后验分布计算得出。

以符合概率论原理的方式纳入模型。事实上，它允许聚类结构从数据中自然显现；此外，利用有关物种的元数据，可以事后检视这一结构，以验证其是否符合关于物种敏感性背后驱动因素的某些科学假设。

BNP模型所诱导的聚类可能与物种的某些特定信息一致，也可能不一致，这可能会对有关物种敏感性决定因素的现有理论提出挑战或提供支持。图3 比较了甲萘威的累积分布函数及其可信带，并展示了估计得到的聚类结构，以及预计与物种敏感性相关的准分类学分组，见de Zwart (2001)。由此得到两个簇：一个主要由甲壳类动物组成，包含除一种以外的所有甲壳类动物；另一个主要由鱼类组成，包含除一种以外的所有鱼类，以及所有软体动物。三种昆虫分布在这两个簇中；唯一的环节动物与甲壳类动物归为一组，而唯一的两栖动物则与鱼类归为一组。因此，对于甲萘威，估计得到的聚类结构似乎与准分类学分组密切相关，这支持了物种敏感性取决于分类学的理论：鱼类形成一个对甲萘威相对耐受的簇，而甲壳类动物则形成一个敏感性更高的簇。然而，并非对所有污染物都能观察到分类学与敏感性之间的这种对应关系；事实上，有些污染物的估计聚类结构与准分类学分组并不一致。对许多污染物的观察发现，鱼类通常倾向于聚为一个簇。这一发现可用于支持减少待测鱼类物种数量的做法，因为只需提高少数代表性物种的权重，便可模拟它们对完整物种敏感性分布的贡献，正如Garnier-Laplace et al. (2006)所做的那样。

4.4 BNP-SSD Shiny应用

上述方法可直接用于`BNPdensity`软件包，但这要求使用者相当熟悉R语言。因此，我们开发了一款名为BNP-SSD的Shiny应用，专为物种敏感性分布问题设计，并基于`BNPdensity`软件包中的函数，可通过`https://alamichl.shinyapps.io/BNP_SSD/`访问。该应用的设计受到`shinyssdtools` 应用的启发，该应用由Dalgarno (2021)开发。

在此应用中，第 2 节所述的 BNP模型将拟合删失或未删失数据。拟合模型前，先清理浓度数据，处理同一物种可能存在多个 CEC 值的情况，再进行 对数变换并中心化缩放。为提高灵活性，部分选项由用户自行设定，例如 MCMC算法的迭代次数。模型拟合完成后，将绘制估计密度以及一些拟合优度图。在另一个面板中，利用分位数的后验分布估计 HC_5 。也可以估计该分布中除 第 5百分位数以外的其他百分位数，并计算该估计的可信带。最后，在最后一个面板中，计算并绘制诱导得到的最优聚类。

5 跨污染物聚类

通过研究数据集中所有污染物的聚类结构，确定类似模式是否普遍存在于污染物中，将会非常有意义。完整的聚类分析需要包含污染物效应的层次模型，这超出了本文的研究范围，将留待未来工作。这里，我们通过对所有污染物分别拟合模型，以一种简单而富有启发性的方式探讨这一问题。因此，我们对各污染物估计得到的聚类结构进行后处理。总体思路是考察所有污染物中每一对物种被归为同一组的频率。由此确定物种的敏感性群落，并将其与准分类学分组进行比较。

我们将研究范围限定为至少由八个物种进行过测试的污染物，这略低于拟合物种敏感性分布时建议的最低阈值（欧洲化学品管理局，2008）。第4.1节介绍了我们的原始数据集，其中包含 3,000 多种污染物。剔除测试物种少于八种的污染物后，仅剩179种。我们首先使用所有这些污染物拟合模型。随后，我们整合各污染物聚类所得的信息，以判断是否存在某些共同模式。为提取各污染物聚类结构中的信息，我们将每个估计得到的聚类转换为一个关联矩阵。将所有关联矩阵依次叠放，便形成一个三维数组（亦称张量），其中每个切片对应一种污染物。一个难点是，不同污染物接受测试的物种集合各不相同，彼此之间可能几乎没有重叠。因此会产生大量缺失值（即尚未进行测试的污染物—物种配对），需要加以处理。

5.1 非负张量分解

我们进行非负三阶张量分解（Cichocki 等，2009），这是一种主成分分析的张量推广方法。它是一种降维技术，可将关联张量分解为多个 R 秩一张量之和。该方法最初在化学计量学领域发展起来，

该技术也已应用于生物统计学、信号处理、语言学和机器学习领域（Gauvin 等人，2014）。该技术还可用于填补缺失值。

我们采用平行因子分析（PARAFAC），也称为典范分解（CANDECOMP）分解；补充材料中的第D 节介绍了该技术的细节以及张量性质的背景。记前文所述数据的张量为$\mathbf{Y}$，则有$\mathbf{Y} \in \mathbb{R}^{n_S \times n_S \times n_C}$ 是一个在前两个维度上对称的张量，其中 n_S 和n_C分别表示物种和污染物的数量。对于某个张量$\hat{}\mathbf{Y} \in \mathbb{R}^{I \times J \times K}$，其一般的 PARAFAC 分解表示为

$$\hat{\mathbf{Y}} = [\![A, B, C]\!] = \sum_{r=1}^{R} a_r \circ b_r \circ c_r,$$

其中 $A = [a_1, \ldots, a_R] \in \mathbb{R}^{I \times R}$、 $B = [b_1, \ldots, b_R] \in \mathbb{R}^{J \times R}$和 $C = [c_1, \ldots, c_R] \in \mathbb{R}^{K \times R}$是三个组分或因子矩阵，而 $\circ$ 表示向量外积。对于所研究的数据，张量$\mathbf{Y}$ 在前两个维度上的对称性意味着 PARAFAC 分解可以简化为$\mathbf{Y} = [\![A, A, C]\!]$，其中 $A = [a_1, \ldots, a_R] \in \mathbb{R}^{n_S \times R}$且 $C = [c_1, \ldots, c_R] \in \mathbb{R}^{n_C \times R}$。请注意，这种分解仅为近似，会产生一定的加性误差$\mathbf{E}$，形式为 $\mathbf{Y} = [\![A, A, C]\!] + \mathbf{E}$。为赋予所得各组分以物理意义，我们采用Xu 和 Yin，2013年提出的 `multiway`R 包（Leeuw，2011）中的非负 PARAFAC 分解。该方法对组分矩阵 $a_{ir} \in \mathbb{R}_+$ 和 $c_{jr} \in \mathbb{R}_+$ 施加非负约束，对所有 $i \in \{1, \ldots, n_S\}$、 $j \in \{1, \ldots, n_C\}$ 以及$r \in \{1, \ldots, R\}$均如此。

一种常用的启发式方法是核心一致性诊断（ R （Bro 和 Kiers，2003），用于确定组分的数量。该诊断需要对缺失值进行插补，以便高效计算。由于我们这类数据中的缺失值数量较多，我们改用交叉验证方法。交叉验证的做法是，移除张量中选定比例的非缺失值，对不同的秩进行分解，然后评估被移除值的重构误差，这是一种 K 折交叉验证。我们使用该张量与原始张量在非缺失值上的弗罗贝尼乌斯距离来衡量重构误差（见图24）。

选定分解的秩后，即可进行分解。三向分解的结果由三个因子矩阵组成，其中两个的维度为$n_S \times R$，另一个的维度为 $n_C \times R$。前两个矩阵表示每个物种属于分解中各个组分的程度，并且按构造二者相等。第三个矩阵则表示每种污染物属于各个组分的程度。为便于解释结果，我们对归属程度设定阈值，并判断每个物种和污染物是否属于某个组分。为此，我们对归属程度向量进行 K均值聚类，以自适应地设定归属程度的阈值。物种和污染物可以被分配到零个或多个组分（见图25 和图26）。

选择这种算法方法，是因为它能够处理缺失数据。在张量分解方面，有多项研究（例如 Rai 等，2014年）提出了一种基于模型的方法，可避免选择分解秩这一问题。然而，在实际应用中，当缺失数据的比例非常高时（如本研究所示），这些方法便会失效。

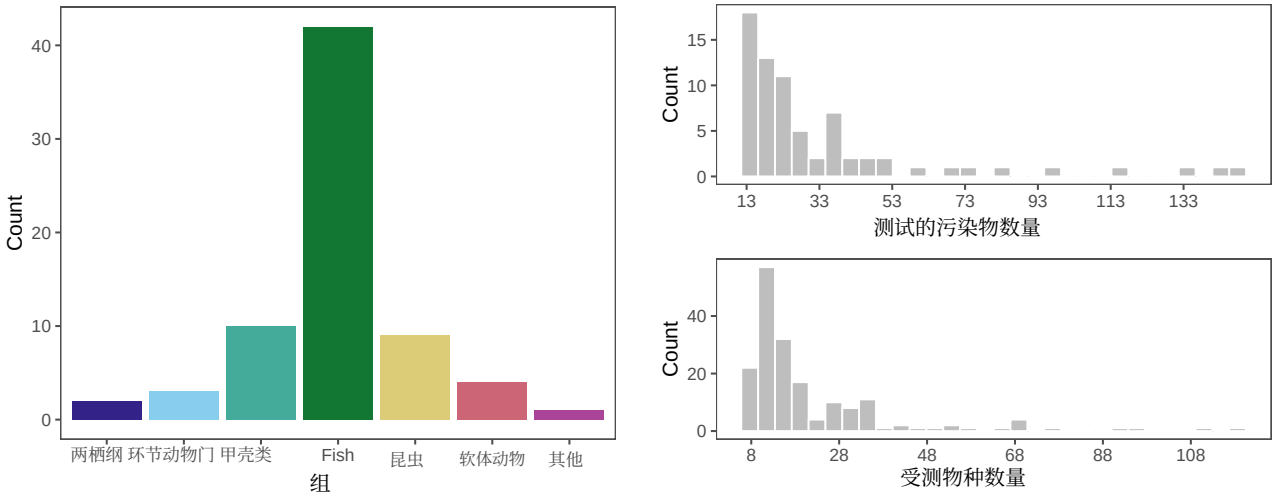

图 4： 左：所研究数据部分中物种的准分类组成。各组根据de Zwart（2001年）中的分类进行定义。右：（上）所研究数据中每个物种受测的污染物数量（每个物种至少受测 13 种污染物）。（下）每种污染物受测的物种数量。

5.2 结果

我们将这一方法应用于第 4.1 节所述数据集的一部分（受限数据集的说明见图 4 ）。为降低缺失数据的比例，我们任意选取了至少测试过 13 种污染物的物种；在这一限制条件下，95 % 的数据缺失。用于进行张量分解的张量维度为$179 \times 71 \times 71$，因此我们有 179 个协同聚类矩阵，每个矩阵包含 71 种物种，并分别对应一种污染物。采用交叉验证方法后，选定的秩为 41 （见图 24）。我们获得了 41个污染物和物种的组分。我们从中展示七个组分，根据图 25和图 26，选取对比度最高的组分，并仅保留按照 K均值聚类标准划分后分离良好的组分。这七个组分将在图5 和图 6中进行分析。

图5 展示了七个组分的准分类组成，以数量（左）和相对比例（右）表示。图 30 和图 29 展示了其他分类等级下的相同结果。我们无法观察到不同组分之间存在明显的组成差异。这表明，准分类似乎并非物种共同聚类的主要驱动因素；换言之，准分类似乎不会显著决定物种敏感性。不过，这一观察结果应结合以下事实加以审视：正如图 4所示，鱼类物种在数据集中占比过高，因此构成了每个聚类的重要部分。

图 6 展示了各污染物在不同组分中的权重。我们可以观察到，只有一种污染物（五氯苯酚）在组分 E 中具有显著权重。在组分 C、D 和 G 中，权重最高的污染物大多是农药。具体而言，组分 C 由两种杀虫剂组成。组分 D 包含两种神经毒性杀虫剂和重铬酸钾，后者是一种氧化剂，用于实验室和工业中的各种反应。组分 G 包含七种杀虫剂，其中包括五种有机磷酸酯类杀虫剂和两种氨基甲酸酯类杀虫剂，以及重铬酸钠，后者

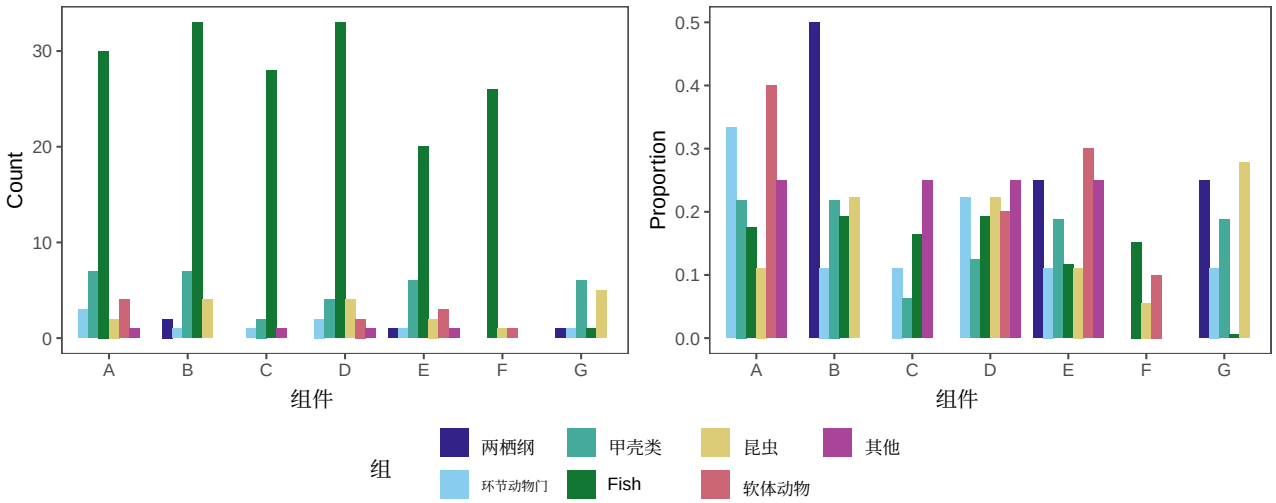

图5：各组分的准分类组成。各组依据de Zwart（2001年）中的分类进行定义。左：各组分中的物种数量。右：各组分中的物种比例与全部数据中物种分布的比较。

主要用作生产其他铬化合物的中间体。在组分 A 中，权重最高的四种污染物均为无机化合物：两种氯化化合物和两种硫酸盐化合物。最后，组分 B 和 F 由性质或用途各异的污染物组成。组分 B 主要包括农药，其中有六种杀虫剂和一种体外寄生虫杀灭剂，此外还有四种具有不同用途的化合物，例如阻燃剂或化学合成中的反应物。组分 F 由两种农药（包括一种杀虫剂和一种鱼类毒杀剂）以及氰化钠组成。氰化钠主要用于黄金提取，但历史上也曾用作杀虫剂。总体而言，这些观察结果表明，各组分与类型相似的污染物相关；换言之，对某一类污染物反应相似的物种，往往也可能对同类型的另一种污染物作出相似反应。这似乎支持这样一种观点：物种对不同污染物的敏感性可能存在相关性；例如，这一观点已被 Awkerman 等人（2008年）

6 讨论与未来研究

我们提出了一种基于贝叶斯非参数混合模型的物种敏感性分布新方法。我们在模拟数据集和真实数据集上与当前的物种敏感性分布模型进行了广泛比较，以展示所提方法的附加价值。当数据集偏离对数正态分布时，BNP-SSD 的表现尤为出色，从而能够充分发挥其在描述数据方面的高度灵活性。与此同时，所提方法相对稳健，似乎不易出现过拟合。BNP-SSD 可被视为介于单组件对数正态物种敏感性分布与组分数量等于物种数量的核密度估计之间的中间模型。事实上，在 RIVM 数据集的所有实际情形中，描述数据所需的簇数量均不超过 3。

BNP-SSD 为风险评估带来了诸多益处：它是一种有效且稳健的标准模型，能够适应众多数据集。因此，BNP-SSD 是一种安全的工具，可

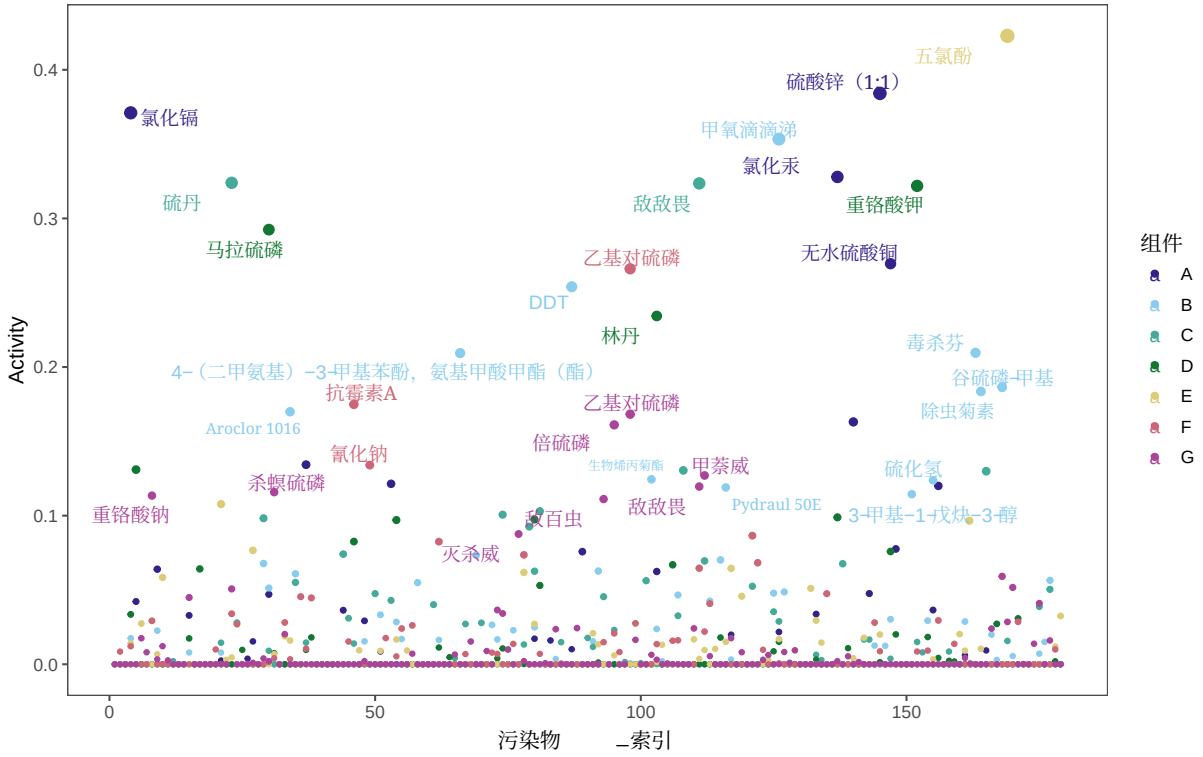

图 6：各组分中每种污染物的权重。点的大小与污染物活性的强度成正比。

去除 SSD（Forbes 和 Calow，2002年）的一项任意参数假设。此外，作为一种贝叶斯方法，它可以直接提供可信区间。

SSD 的传统方法是将污染物分别进行考量。在数据稀缺的背景下，减少动物试验的推动进一步加剧了这一问题，因此，充分利用生态毒性测试的长期积累，从涉及不同污染物的实验中借鉴信息，将是理想之举。大型数据库已经可供使用。我们可以期待，围绕化学品监管透明度的持续讨论，甚至会推动数据更大范围地向公众开放。因此，开发能够利用现有关于物种对其他污染物敏感性的全部信息的模型，正逢其时。从本质上说，这正是 Awkerman 等人（2008）和 Craig（2013）提出的方案：利用分类学信息预测未知物种对某种污染物的敏感性。BNP-SSD 方法的一个重要副产品，是为后续聚类分析提供了有趣的机会。我们展示了一个使用非负张量分解进行后处理分析的实例，总结了关于179种污染物的洞见；分析结果表明，物种敏感性可能会基于污染物类型呈现某些规律。然而，我们没有观察到跨污染物的分类学与物种敏感性之间存在强关联。因此，我们可以假设，除分类学结构之外，可能还存在其他与物种敏感性相关的结构，例如生态位。

这一思路促使我们自然地拓展当前研究：可利用依赖型贝叶斯非参数模型，将贝叶斯非参数物种敏感性分布方法扩展为能够刻画污染物间变异的方法。这一点尤为重要，因为某些污染物可能具有非常相似的毒性，这通常是因为它们属于同一化学类别。现有多种依赖型贝叶斯非参数模型

采用不同机制来引入依赖关系，包括层级（Teh 等人，2006年；Camerlenghi 等人，2019b）、加性（Müller 等人，2004年； Lijoi 等人，2014）、嵌套（Rodríguez 等人，2008年；Camerlenghi 等人，2019a）或这些机制的组合。近期，Franzolini 等人（2025）在一个通用框架中统一了这些模型。未来的研究将深入探讨哪种依赖结构及相应的信息共享方式最适用于物种敏感性分布的情境。

此外，正如在研究物种聚类时所讨论的，组既可以根据已知的污染物化学性质来指定，也可以通过贝叶斯非参数方法灵活地学习得到。这将为系统研究性质可能截然不同的污染物之间的相似性打开大门。此外，我们也希望不再仅用单一值概括物种敏感性。生态毒理学测试通常通过拟合剂量-反应模型来分析，该模型描述物种对污染物的多种响应特征，例如从暴露到产生效应之间的时间。使用剂量-反应模型中的所有参数，就意味着要在更高维的空间中进行聚类，从而增强组间的区分能力，并最终得到更有意义的簇。

致谢

L. Alamichel 部分获得 LabEx PERSYVAL-Lab（ANR-11-LABX-0025-01）的资助，该项目由法国“未来投资计划”资助。J. Arbel 部分获得法国国家研究署 ANR-21-JSTM-0001 项目的资助。I. Prünster 部分获得欧盟—下一代欧盟 PRIN-PNRR（项目 P2022H5WZ9）的资助。

参考文献

Aldenberg, T. 和 Jaworska, J. S.（2000年）正态物种敏感性分布的有害浓度及受影响比例的不确定性。《生态毒理学与环境安全》，**46**，1–18。（引用页码：3，5，6和8。）

Aldenberg, T.、Jaworska, J. S. 和 Traas, T. P.（2002年）正态物种敏感性分布与概率生态风险评估。载于《生态毒理学中的物种敏感性分布》（编者：L. Posthuma、G. I. Suter 和 T. P. Traas），49–102。佛罗里达州博卡拉顿：Lewis 出版社。（引用页码2。）

Aldenberg, T. 和 Slob, W.（1993）基于逻辑分布的 NOEC 毒性数据计算有害浓度的置信限。《生态毒理学与环境安全》，**25**，48–63。（引用页码2。）

ANZECC（2000年）《澳大利亚和新西兰淡水及海水水质指南》。技术报告，澳大利亚和新西兰环境与保护委员会、澳大利亚和新西兰农业与资源管理委员会，澳大利亚堪培拉。（引文见第2页。）

Arbel, J., KonKamKing, G., Lijoi, A., Nieto-Barajas, L.E. 和 Prünster, I.（2021）BNPdensity：R 中的贝叶斯非参数混合建模。《澳大利亚与新西兰统计学杂志》，**63**，542–564。（引文见第4页。）

Arbel, J. 和 Prünster, I.（2017）矩匹配 Ferguson–Klass 算法。《统计与计算》，27，3–17。（引文见第 6 页。）Awkerman, J. A., Raimondo, S. 和 Barron, M. G.（2008年）利用物种间毒性相关模型为野生动物构建物种敏感性分布。《环境科学与技术》，42，3447–3452。（引文见第 3、16 和 17 页。）Barrios, E., Lijoi, A., Nieto-Barajas, L. E. 和 Prünster, I.（2013年）归一化随机测度混合模型。《统计科学》，28，313–334。（引文见第 3、4 和 5 页。）Binder, D. A.（1978）贝叶斯聚类分析。《生物计量学》，65，31–38。（引文见第 8 页。）Bro, R. 和 Kiers, H. A. L.（2003）一种确定 PARAFAC 模型中组分数量的新型高效方法。《化学计量学杂志》，17，274–286。（引文见第 14 页。）Camerlenghi, F., Dunson, D. B., Lijoi, A., Prünster, I. 和 Rodríguez, A.（2019a）潜在嵌套非参数先验（附讨论）。《贝叶斯分析》，14，1303–1356。（引文见第 18 页。）Camerlenghi, F., Lijoi, A., Orbanz, P. 和 Prünster, I.（2019b）分层过程的分布理论。The Annals of Statistics，47，67–92。（引文见第 18 页。）

CCME（2007）保护水生生物的水质指南推导规程。载于加拿大环境质量指南，编号 Ccme 1991。温尼伯：加拿大环境部长理事会。（引用于第2页。）

Chen, L.（2004年）物种敏感性分布第5百分位数的一种保守非参数估计量。《统计规划与推断杂志》，123，243–258。（引用于第2页。）Cichocki, A.、Zdunek, R.、Phan, A. H.与Amari, S.-i.（2009年）非负矩阵与张量分解：在探索性多路数据分析和盲源分离中的应用。约翰·威立父子公司。（引用于第13页。）Craig, P. S.（2013年）探索利用物种敏感性分布确定工业化学品预测无效应浓度的新方法：致项目指导小组的最终报告。技术报告。（引用于第3、11和17页。）Craig, P. S.、Hickey, G. L.、Luttik, R.与Hart, A.（2012年）概率生态毒理风险评估中的物种不可交换性。《英国皇家统计学会杂志 A辑：社会统计》，175，243–262。（引用于第3页。）Dahl, D. B.（2006年）通过狄利克雷过程混合模型对表达数据进行基于模型的聚类。《基因表达与蛋白质组学的贝叶斯推断》，201–218。（引用于第7页。）Dahl, D. B.、Johnson, D. J.与Müller, P.（2022年）贝叶斯聚类的搜索算法与损失函数。Journal of Computational and Graphical Statistics，31，1189–1201。（引用于第8页。）Dalgarno, S.（2021）shinyssdtools：用于拟合物种敏感性分布（物种敏感性分布）的网络应用程序。《开源软件杂志》，6，2848。（引用于第13页。）

欧洲化学品管理局（2008年）环境剂量浓度-效应关系的表征。载于《信息要求与化学品安全评估指南》，编号May，第R.10章。赫尔辛基：欧洲化学品管理局。（引用于第2、3、11和13页。）Ferguson, T.与Klass, M.（1972年）不含高斯分量的独立增量过程的一种表示。《数学统计年刊》，43，1634–1643。（引用于第4页。）Forbes, V. E.与Calow, P.（2002年）重新审视物种敏感性分布：一项批判性评价。《人类与生态风险评估》，8，473–492。（引用于第2和17页。）Franzolini, B.、Lijoi, A.、Prünster, I.与Rebaudo, G.（2025年）多变量物种抽样模型。网址：https://arxiv.org/abs/2503.24004。（引用于第18页。）Garnier-Laplace, J.、Della-Vedova, C.、Gilbin, R.、Copplestone, D.、Hingston, J.与Ciffroy, P.（2006年）受放射性物质暴露的淡水与陆地生态系统预测无效应浓度的首次推导。《环境科学与技术》，40，6498–6505。（引用于第12页。）Gauvin, L.、Panisson, A.与Cattuto, C.（2014年）检测时序网络的群落结构与活动模式：一种非负张量分解方法。PLoS ONE，9。（引用于第14页。）Gelfand, A. E.（1996年）使用基于抽样的方法进行模型确定。《实践中的马尔可夫链蒙特卡罗》，145–161。（引用于第7页。）Grist, E. P. M.、Leung, K. M. Y.、Wheeler, J. R.与Crane, M.（2002年）水生生物群落有害浓度阈值的改进自助法估计。《环境毒理学与化学》，21，1515–1524。（引用于第2页。）He, W.、Qin, N.、Kong, X.、Liu, W.、Wu, W.、He, Q.、Yang, C.、Jiang, Y.、Wang, Q.、Yang, B.与Xu, F.（2014年）使用贝叶斯 Matbugs 计算器（BMC）对京津冀环渤海地区水体中典型有毒污染物进行生态风险评估与优先级排序。《生态指标》，45，209–218。（引用于第2页。）Hickey, G. L.、Craig, P. S.、Luttik, R.与de Zwart, D.（2012年）生态毒性数据试验间变异的量化及其在物种敏感性分布中的应用。《环境毒理学与化学》，31，1903–1910。（引用于第10页。）Jagoe, R. H.与Newman, M. C.（1997年）群落无可观察效应浓度的自助法估计。《生态毒理学》，6，293–306。（引用于第2页。）Jones, D. S.、Barnthouse, L. W.、Suter II, G. W.、Efroymson, R. A.、Field, J. M.与Beauchamp, J. J.（1999年）大型河流—水库中的生态风险评估：3. 底栖无脊椎动物。《环境毒理学与化学》，18，599–609。（引用于第2页。）Jordan, M. I.（2010年）分层模型、嵌套模型与完全随机测度。载于《统计决策与贝叶斯分析前沿：纪念James O. Berger》，207–218。纽约：施普林格。（引用于第4页。）

Kefford, B. J., Hickey, G. L., Gasith, A., Ben-David, E., Dunlop, J. E., Palmer, C. G., Allan, K., Choy, S. C. and Piscart, C. (2012) 河流大型无脊椎动物盐度敏感性的全球尺度变异：澳大利亚东部、法国、以色列和南非。PLoS ONE, 7, e35224.（引用于第3页和第11页。）Kingman, J. (1975) 随机离散分布。英国皇家统计学会会刊B辑, 37, 1–22.（引用于第5页。）Kingman, J. F. C. (1967) 完全随机测度。Pacific Journal Of Mathematics, 21, 59–78.（引用于第4页。）Kon Kam King, G., Veber, P., Charles, S. and Delignette-Muller, M. L. (2014) MOSAIC_SSD：一种通过最大似然将删失数据纳入物种敏感性分布的新网络工具。《环境毒理学与化学》, 33, 2133–2139.（引用于第3、5和11页。）Kooijman, S. (1987) 考虑物种间敏感性差异的LC50值安全系数。Water Research, 21, 269–276.（引用于第2页。）Lau, J. W. and Green, P. J. (2007) 基于贝叶斯模型的聚类程序。Journal of Computational and Graphical Statistics, 16, 526–558.（引用于第7页。）Leeuw, J. d. (2011) Multiway软件包。（引用于第14页。）Lijoi, A., Mena, R. H. and Prünster, I. (2007) 控制贝叶斯非参数混合模型中的强化。英国皇家统计学会会刊B辑：统计方法学, 69, 715–740.（引用于第5页。）Lijoi, A., Nipoti, B. and Prünster, I. (2014) 基于相依归一化完全随机测度的贝叶斯推断。Bernoulli, 20, 1260–1291.（引用于第18页。）Lijoi, A. and Prünster, I. (2010) 超越狄利克雷过程的模型。载于贝叶斯非参数方法（编：N. L. Hjort, C. C. Holmes, P. Müller and S. G. Walker），第28卷，80。剑桥大学出版社，剑桥。（引用于第4页和第5页。）Liu, Y., Wu, F., Mu, Y., Feng, C., Fang, Y., Chen, L. and Giesy, J. P. (2014) 在中国制定水质标准：针对金属和类金属建立物种敏感性分布的方法。载于Reviews of Environmental Contamination and Toxicology卷，35–57。施普林格。（引用于第2页。）Lo, A. Y. (1984) 论一类贝叶斯非参数估计：I. 密度估计。The Annals of Statistics, 12, 351–357.（引用于第4页。）Meilă, M. (2007) 比较聚类——一种基于信息的距离。Journal of Multivariate Analysis, 98, 873–895.（引用于第8页。）Müller, P., Quintana, F. and Rosner, G. L. (2004) 一种合并相关非参数贝叶斯模型推断的方法。J. R. Stat. Soc. Ser. B, 66, 735–749.（引用于第18页。）Posthuma, L., Suter II, G. W. and Trass, P. T. (2002) 生态毒理学中的物种敏感性分布。CRC 出版社。（引用于第2页。）

Rai, P., Wang, Y., Guo, S., Chen, G., Dunson, D. and Carin, L. (2014) 不完整多路张量的可扩展贝叶斯低秩分解。载于第31届国际机器学习会议论文集，1800–1808。PMLR。（引用于第14页。）Rastelli, R. and Friel, N. (2018) 潜变量聚类模型的最优贝叶斯估计量。《统计与计算》, 28, 1169–1186.（引用于第8页。）Regazzini, E., Lijoi, A. and Prünster, I. (2003) 具有独立增量的归一化随机测度之均值的分布结果。《统计年刊》, 31, 560–585.（引用于第4页。）Rodríguez, A., Dunson, D. B. and Gelfand, A. E. (2008) 嵌套狄利克雷过程。《美国统计学会会刊》, 103, 1131–1154.（引用于第18页。）Roux, D. J., Jooste, S. H. J. and MacKay, H. M. (1996) 保护南非淡水生态系统的特定物质水质标准：部分无机有毒物质的推导方法与初步结果。《南非科学杂志》, 92, 198–206.（引用于第2页。）Shao, Q. (2000年) 基于NOEC毒性数据的有害浓度估计：一种替代方法。《环境计量学》, 11, 583–595.（引用于第2页。）Silverman, B. W. (1986) 统计与数据分析的密度估计，第26卷。CRC 出版社。（引用于第5页。）Suter II, G. W., Barnthouse, L. W., Efroymson, R. A. and Jager, H. (1999年) 大型河流—水库中的生态风险评估：2. 鱼类群落。《环境毒理学与化学》, 18, 589–598.（引用于第2页。）Teh, Y., Jordan, M., Beal, M. J. and Blei, D. M. (2006) 层次狄利克雷过程。J. Am. Stat. Assoc., 101, 1566–1581.（引用于第18页。）美国环保署 (1998) 生态风险评估指南。技术报告，美国环境保护局，华盛顿特区。（引用于第2页。）Van Der Hoeven, N. (2001年) 在不对分布作任何假设的情况下估计物种敏感性分布的第5百分位数。《生态毒理学》, 10, 25–34.（引用于第2页。）Van Straalen, N. M. (2002年) 应用于锌水生生态风险评估的物种敏感性分布阈值模型。《环境毒理学与药理学》, 11, 167–172.（引用于第2页。）Wade, S. and Ghahramani, Z. (2018) 贝叶斯聚类分析：点估计和可信球。《贝叶斯分析》, 13, 559–626.（引用于第8页。）Wagner, C. and Lokke, H. (1991) 由NOEC毒性数据估计生态毒理学保护水平。Water Research, 25, 1237–1242.（引用于第2页。）Wang, B., Yu, G., Huang, J. and Hu, H. (2008年) 采用五种统计方法建立有机氯农药的物种敏感性分布并估计HC(5)。《生态毒理学》, 17, 716–724.（引用于第2页。）

Wang, Y., Wu, F., Giesy, J. P., Feng, C., Liu, Y., Qin, N. 和 Zhao, Y.（2015年）用于制定金属水质标准的物种敏感性分布非参数核密度估计。环境科学与污染研究，**22**，13980–13989。（引文见第2、3、5和第8页。）

Xing, L., Liu, H., Zhang, X., Hecker, M., Giesy, J. P. 和 Yu, H.（2014）用于保护水生生物的淡水水质标准制定中统计方法的比较。环境科学与污染研究，**21**，159–167。（引文见第2页。）

Xu, F.-L., Li, Y.-L., Wang, Y., He, W., Kong, X.-Z., Qin, N., Liu, W.-X., Wu, W.-J. 和 Jorgensen, S. E.（2015年）生态风险评估中物种敏感性分布（SSD）模型开发与应用的关键问题。《生态指标》，**54**，227–237。（引文见第2页。）

Xu, Y. 和 Yin, W.（2013年）一种用于正则化多凸优化的块坐标下降法及其在非负张量分解与补全中的应用。*SIAM* 成像科学期刊，**6**，1758–1789。（引自第14页。）

Zajdlik, B. A.、Dixon, D. G. 和 Stephenson, G.（2009年）利用多模态物种敏感性分布制定环境污染物水质指南：阿特拉津案例研究。《人类与生态风险评估》，**15**，554–564。（引自第3页。）

Zhao, J. 和 Chen, B.（2016）毒死蜱对水生生物的物种敏感性分布：模型选择与样本量。《生态毒理学与环境安全》，**125**，161–9。（引自第2页。）

de Zwart, D.（2001年）水生物种物种敏感性分布中观察到的规律。载于《生态毒理学中的物种敏感性分布》。CRC 出版社。（引自第8、10、11、12、15和16页。）

补充材料

本补充材料安排如下：第A 节提供真实数据的补充结果，第B 节展示真实数据结果的收敛诊断，第C节讨论模型参数的敏感性分析，第D 节介绍非负张量分解的细节，第E 节展示与主文档第5节相关的补充图。

A 真实数据结果

本节展示第4.2节所述的结果。我们给出三种模型（贝叶斯非参数、核密度估计和正态）在真实数据上的比较。我们考虑三类污染物：拥有大型数据集的污染物（包含超过60个值）、中型数据集（约25个值）以及小型数据集（略多于10个值）。三种模型分别拟合于每个数据集，我们研究了HC_5 的估计值及其可信区间、留一法误差，以及估计密度相对于直方图的形状。同一批数据集的删失版本也用贝叶斯非参数模型和正态模型进行了研究。

A.1 非删失数据的模型比较

我们给出这些数据集非删失版本的结果。

A.1.1 大型未删失数据集

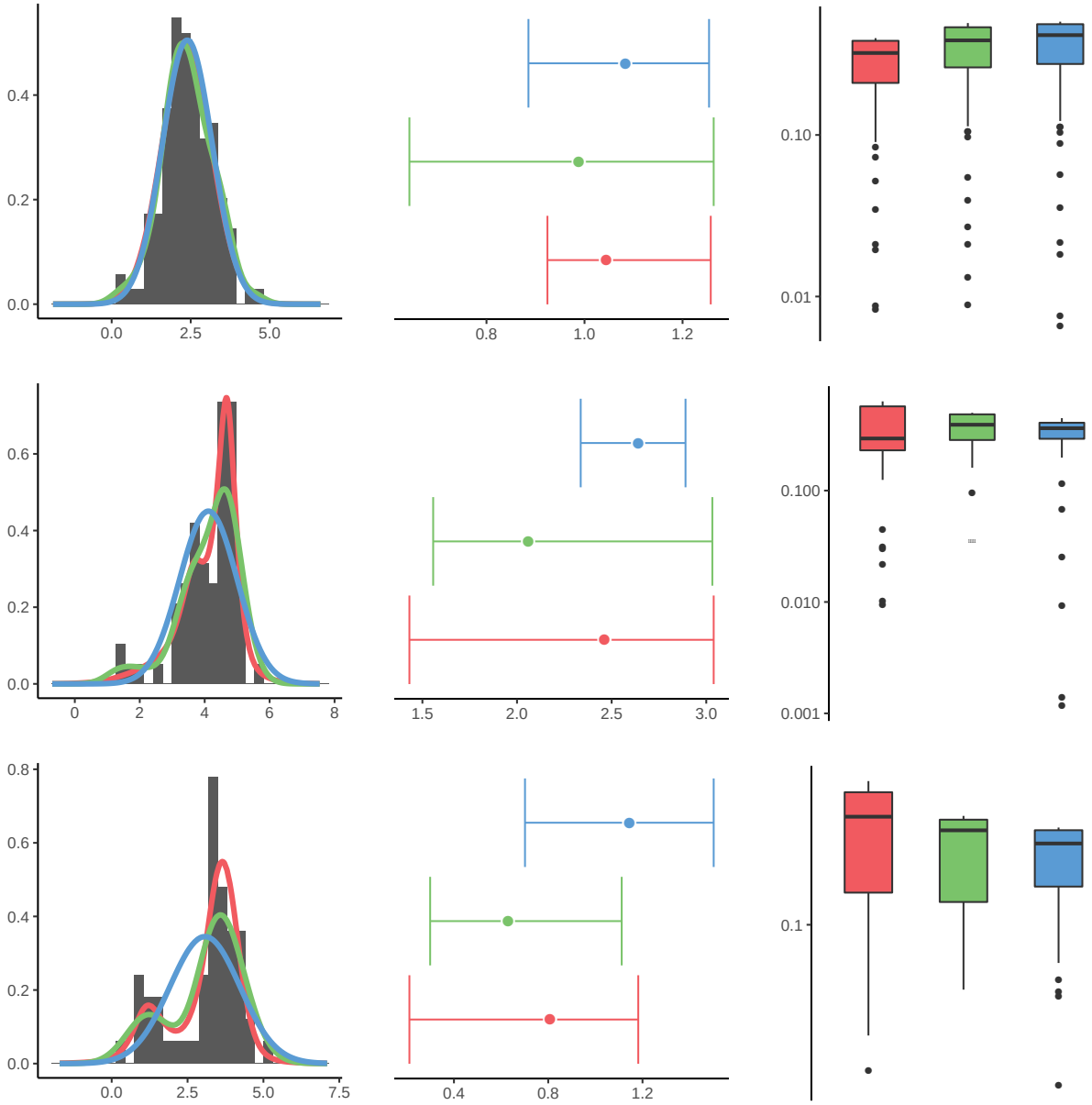

图 7： 密度估计，HC_5 以及三个大型非删失数据集的 CPO。红色（ ）表示 BNP 模型，蓝色（ ）表示正态模型，绿色（ ）表示 KDE 模型。从上到下依次为：氯化镉、重铬酸钾和甲萘威。

A.1.2 中等规模未删失数据集

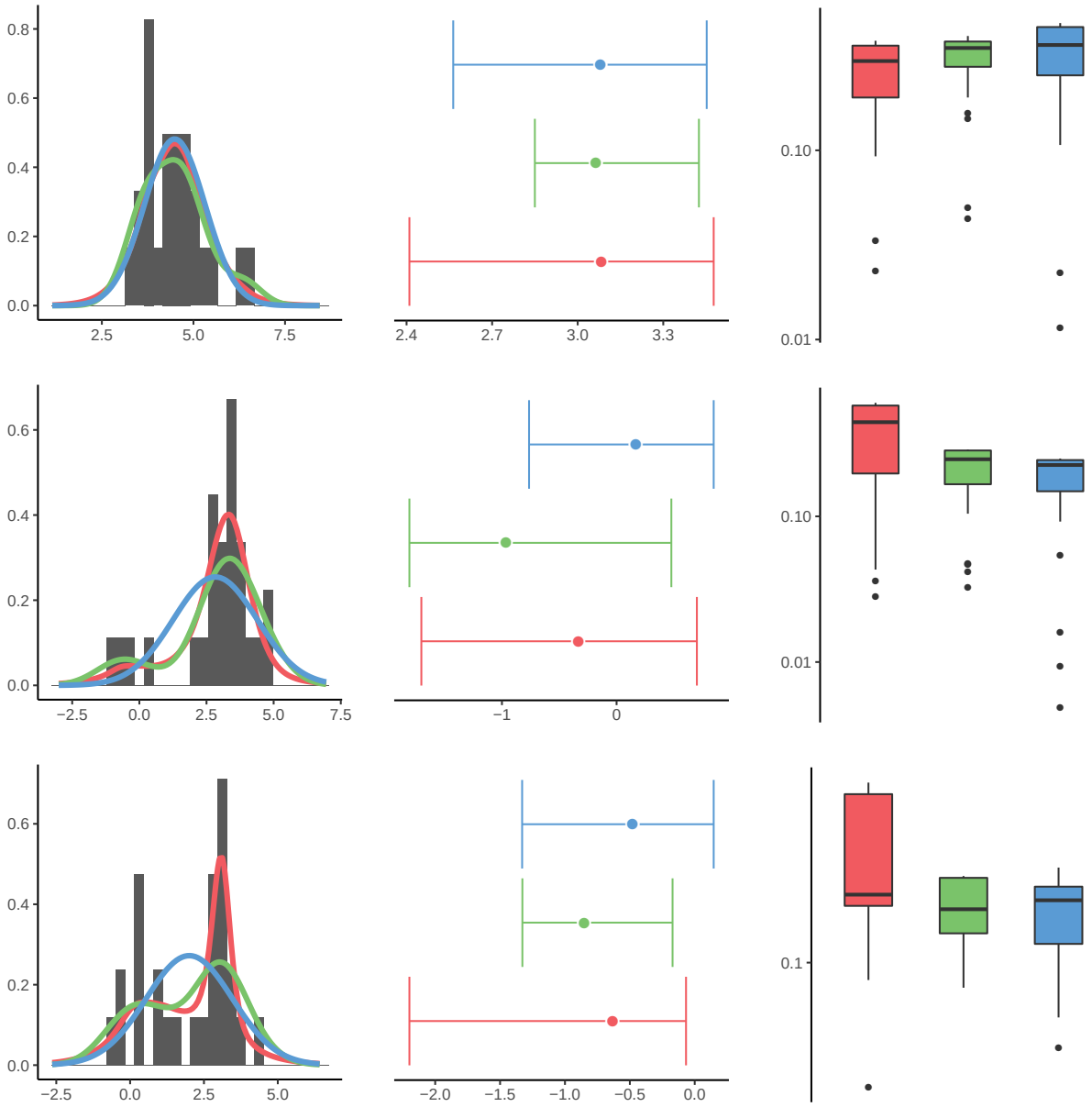

图 8： 密度估计、HC_5 以及三个中等规模未删失数据集的 CPO。红色（ ）表示贝叶斯非参数模型，蓝色（ ）表示正态模型，绿色（ ）表示核密度估计模型。从上到下：2,4-D 酸、敌百虫、对硫磷。

A.1.3 小型未删失数据集

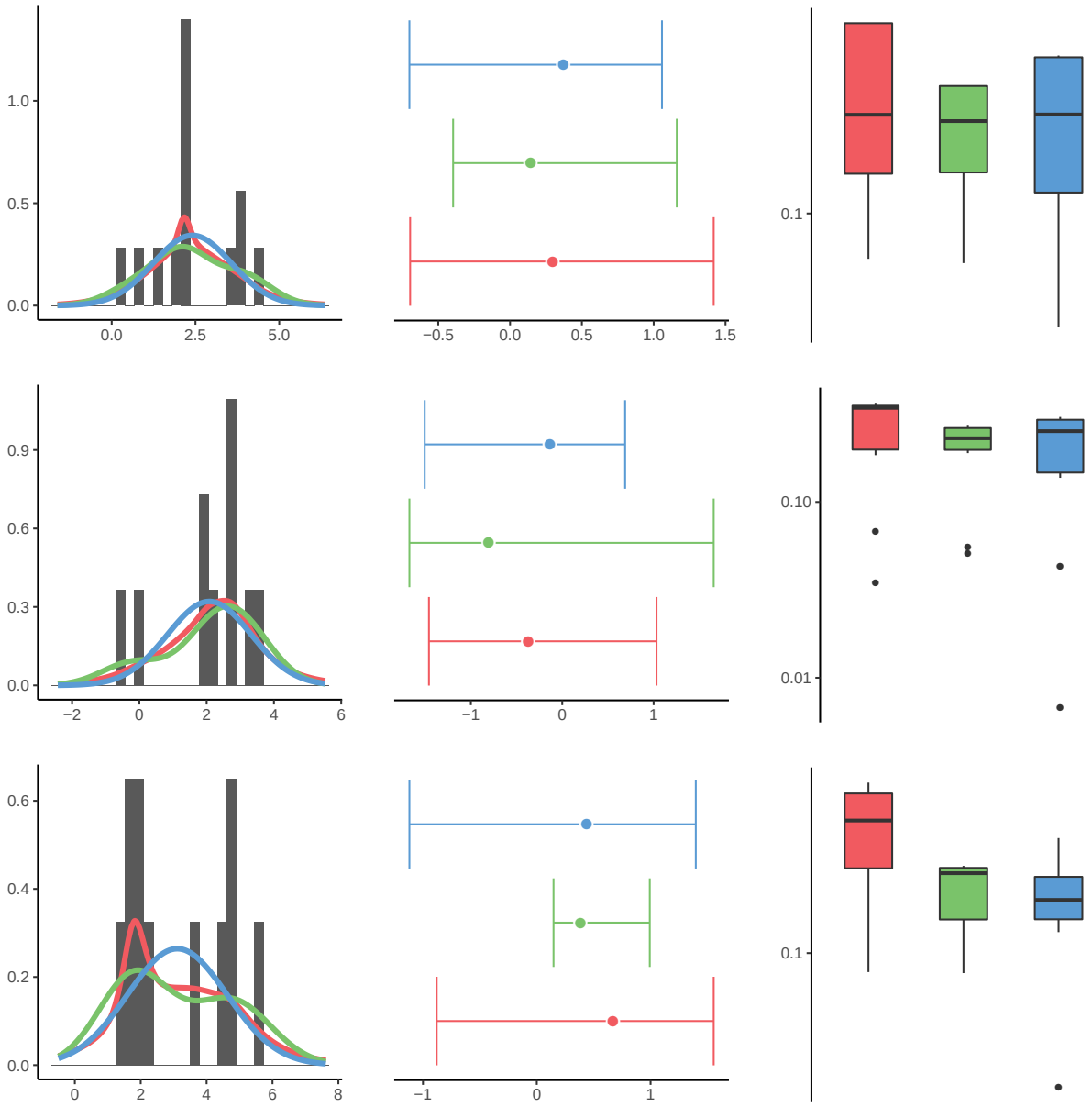

图 9：密度估计，HC_5 以及三个小型未删失数据集的 CPO。红色（ ）表示 BNP 模型，蓝色（ ）表示正态模型，绿色（ ）表示 KDE 模型。从上到下依次为：伏蚜磷、二溴磷、重铬酸钠。

A.2 删失数据集上的模型比较

这里展示数据集删失版本的结果。由于目前尚无适用于删失数据的核密度估计器扩展（见第 2.1 节），我们只比较正态模型和贝叶斯非参数模型。

A.2.1 大型删失数据集

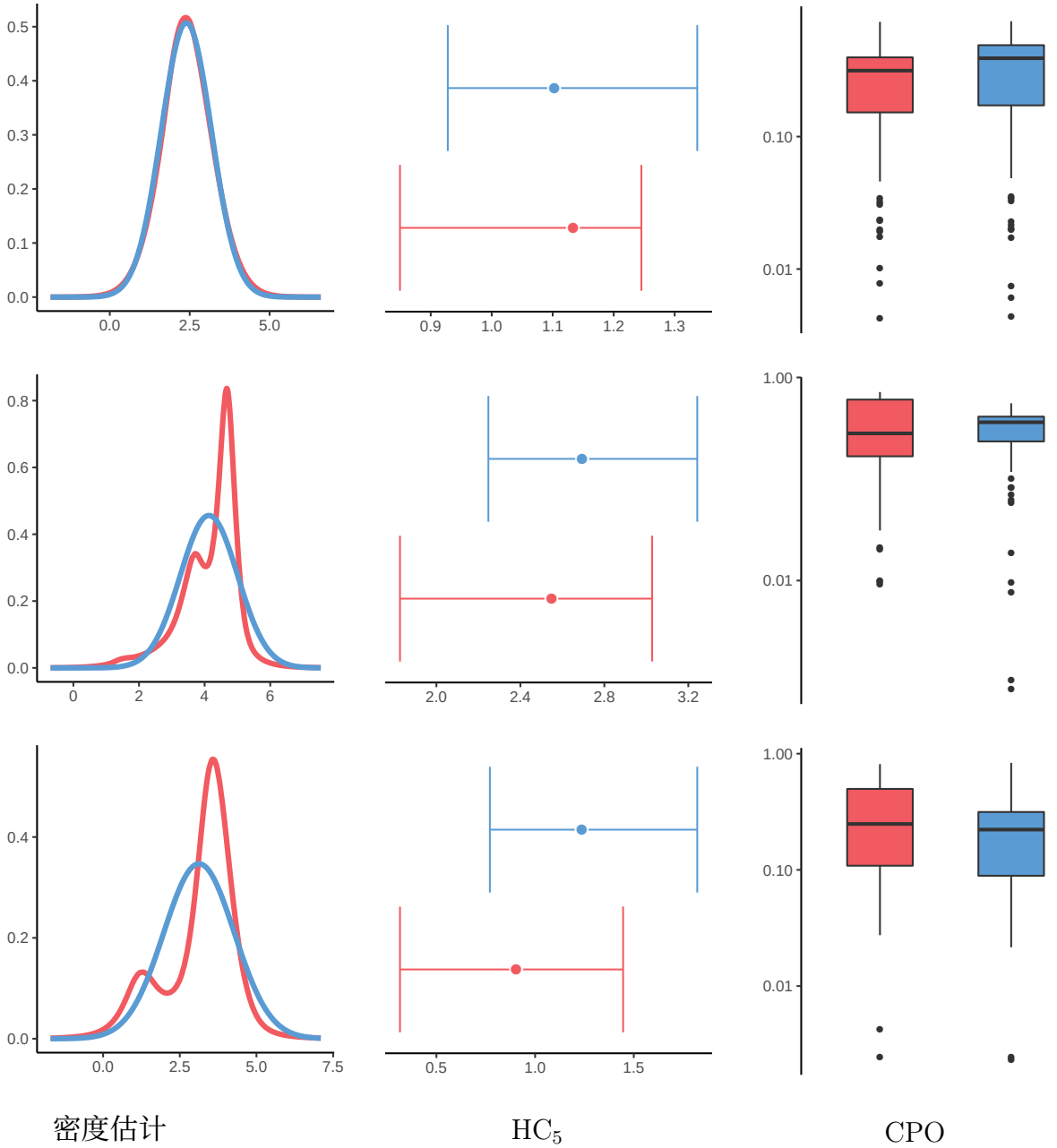

图 10： 三个大型删失数据集的密度估计、HC_5 和 CPO。贝叶斯非参数模型用红色（ ）表示，正态模型用蓝色（ ）表示（尚未针对删失数据实现核密度估计）。从上到下依次为：氯化镉、重铬酸钾和甲萘威。

A.2.2 中度删失数据集

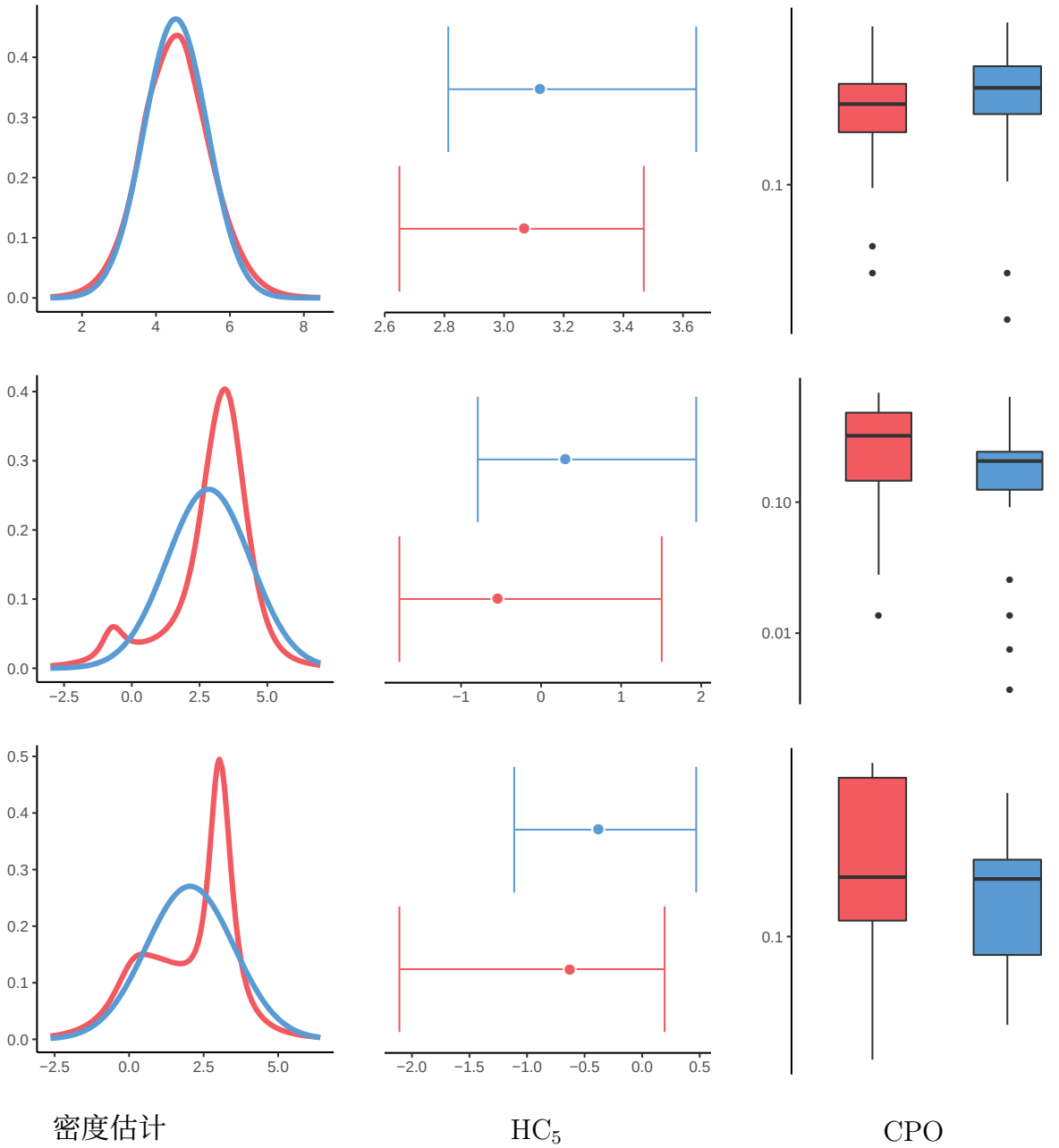

图 11： 密度估计，HC_5 以及三个中度删失数据集的 CPO。红色（ ）表示 BNP 模型，蓝色（ ）表示正态模型（尚未针对删失数据实现核密度估计）。从上到下依次为：2,4-D 酸、敌百虫、对硫磷。

A.2.3 小型删失数据集

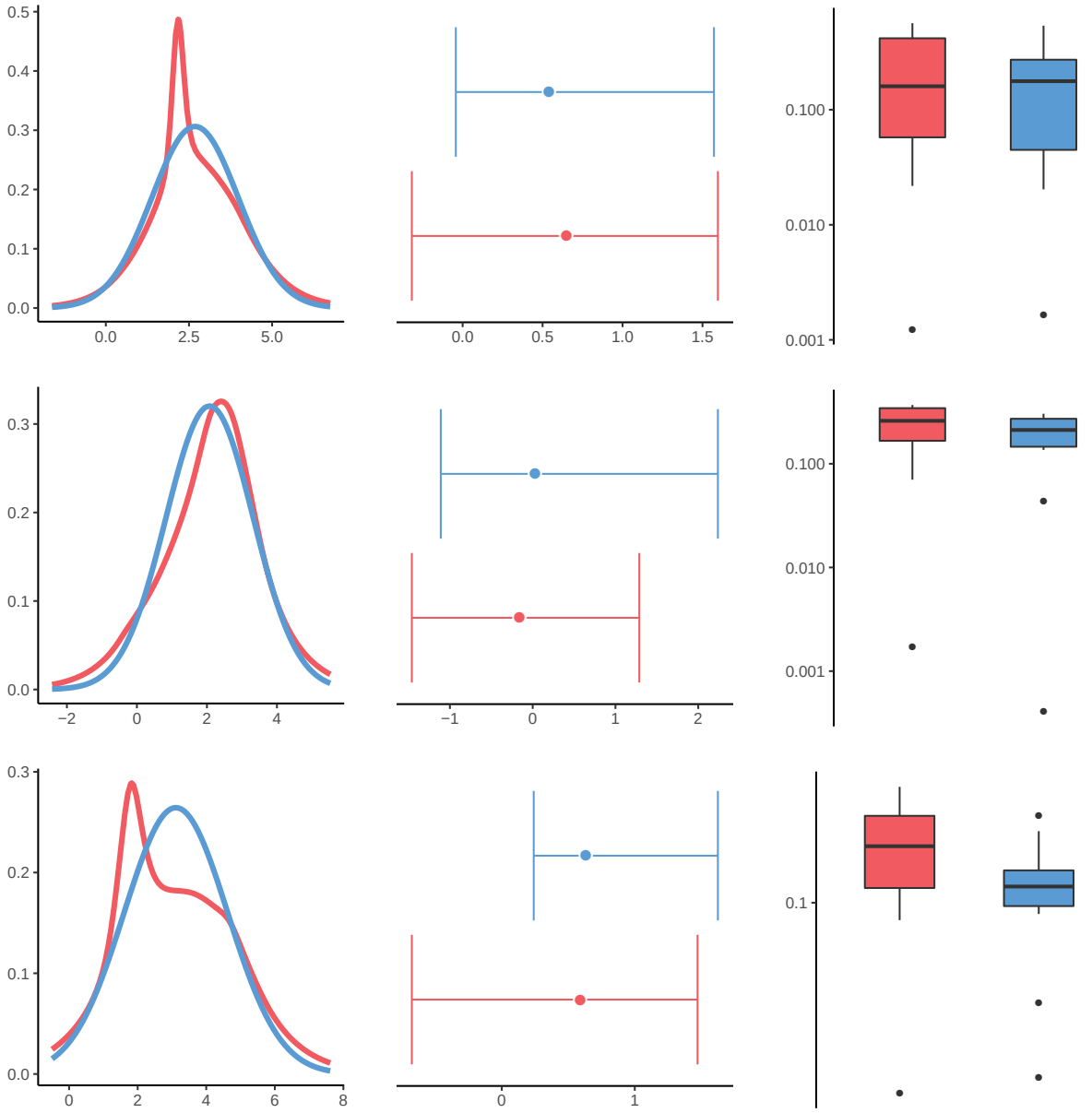

图12： 密度估计，HC_5 以及三个小型删失数据集的CPO。红色（）表示BNP模型，蓝色（）表示正态模型（核密度估计未针对删失数据实现）。从上到下：伏蚜磷、二溴磷、重铬酸钠。

B 真实数据分析的收敛诊断A

以下结果基于未删失数据集得出。对于删失版本的数据集，轨迹图也类似。

B.1 大型数据集

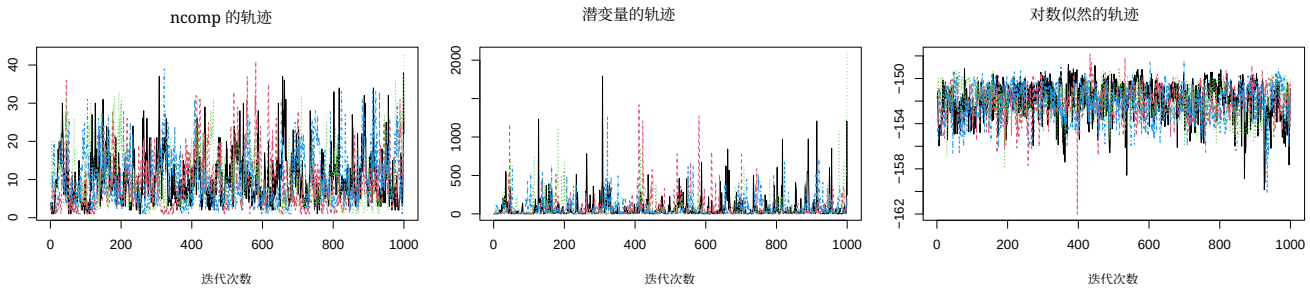

图 13： 使用 BNPdensity 和 coda 软件包得到的氯化镉轨迹图 e.

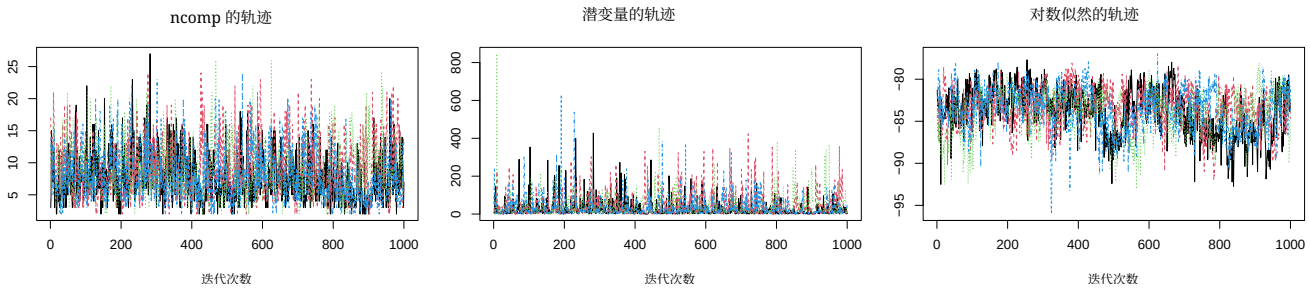

图 14： 重铬酸钾的轨迹图。

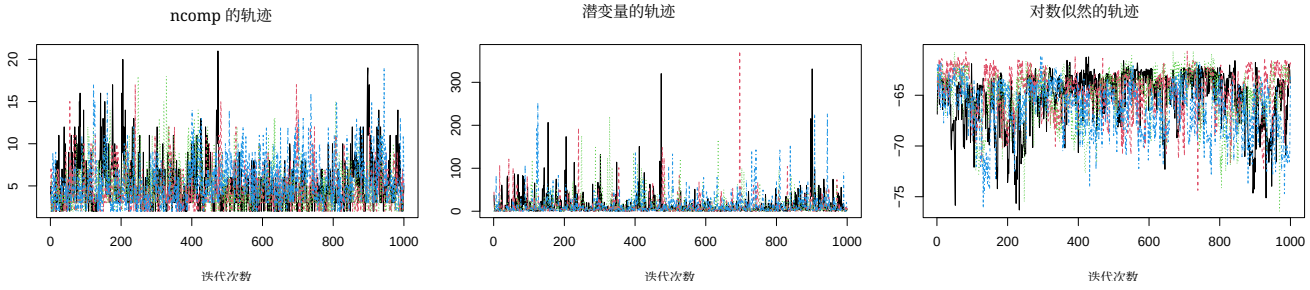

图 15： 甲萘威的轨迹图。

	有效样本量（成分数）	有效样本量（潜在_变量）	有效样本量（对数_似然）
氯化镉	171	341	43
重铬酸钾	71	121	59
甲萘威	242	400	37

表 1：不同数据集参数的有效样本量。

B.2 中型数据集

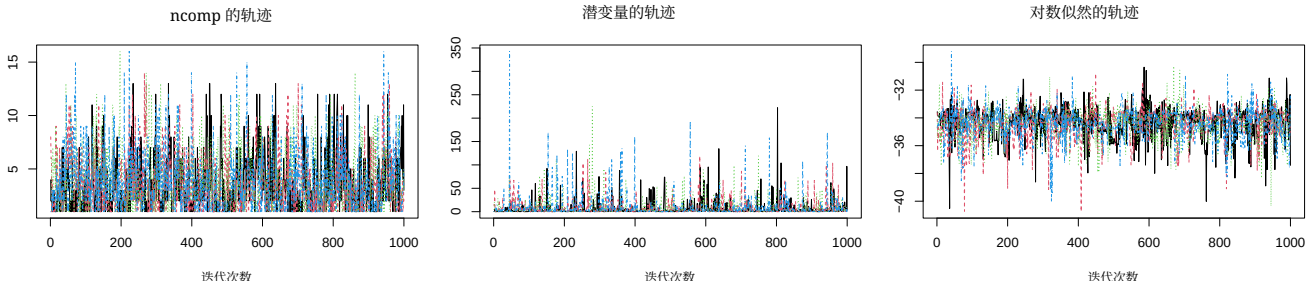

图 16：2,4-D 酸的轨迹图。

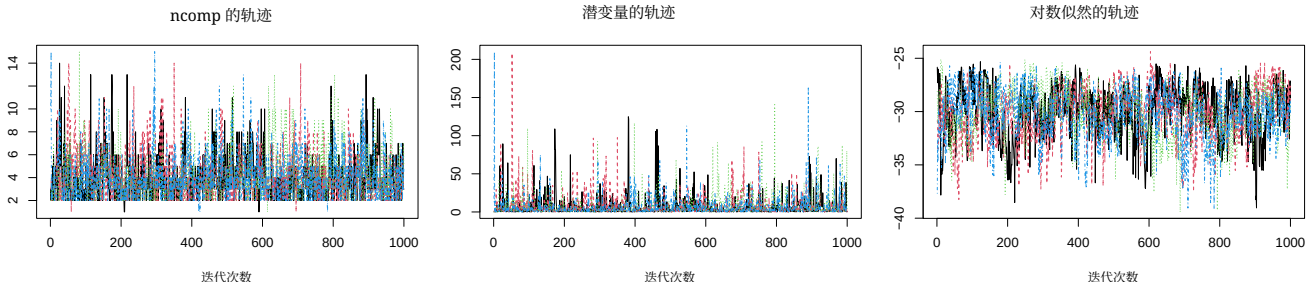

图 17：敌百虫的轨迹图。

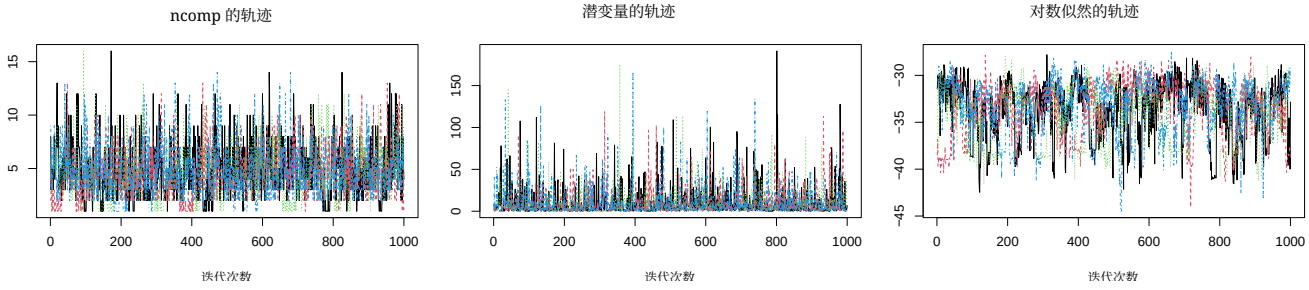

图 18：乙基对硫磷的轨迹图。

	有效样本量（成分数）	有效样本量（潜在_变量）	有效样本量（对数_似然）
2,4-D酸	201	316	213
敌百虫	276	404	37
乙基对硫磷	190	501	41

表 2： 不同数据集参数的有效样本量。

B.3 小型数据集

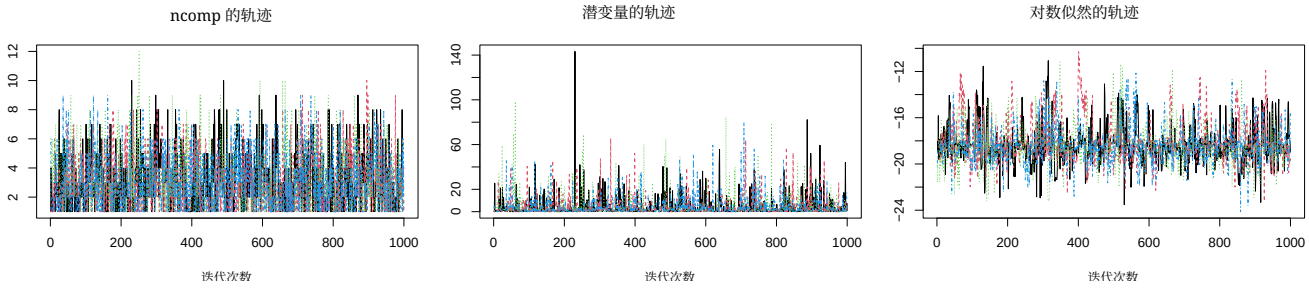

图19： 伏蚜磷的轨迹图。

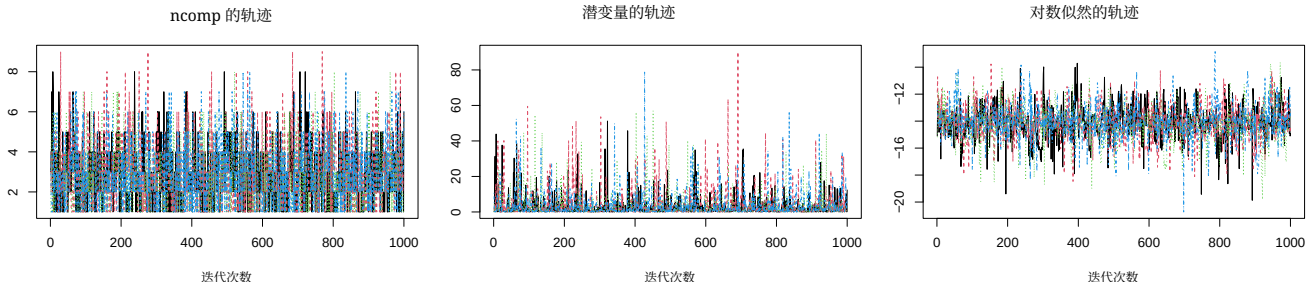

图20： 二溴磷的轨迹图。

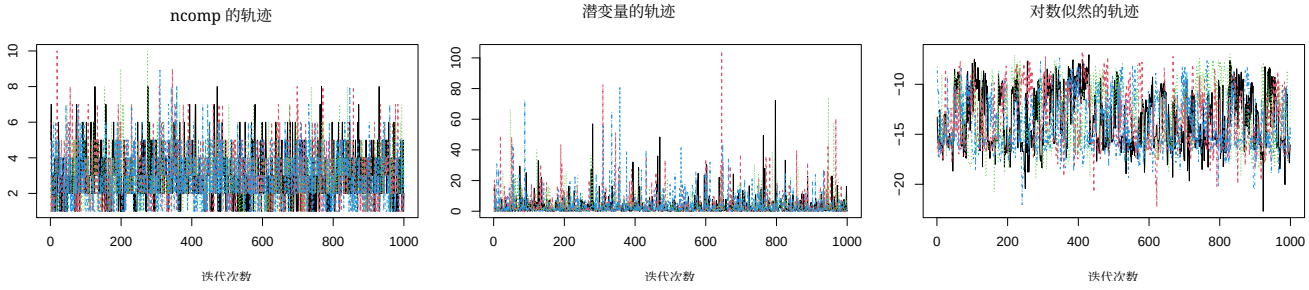

图21： 重铬酸钠的轨迹图。

	有效样本量（成分数）	有效样本量（潜在_变量）	有效样本量（对数_似然）
伏蚜磷	286	585	153
二溴磷	433	624	318
重铬酸钠	350	590	77

表3： 不同数据集参数的有效样本量。

C 模型的敏感性分析

C.1 参数 γ

我们开展了一项实验，以检验第 2 节所述模型对参数 γ选择的稳健性。我们使用模型对一个玩具数据集的密度进行估计。该玩具数据集名为 acidity，取自 `BNPdensity`包。我们测试了参数 $\gamma \in (0,1)$的不同取值。的这些取值会导致不同的先验行为。结果汇总于图 22中。从图中可以看出，估计得到的不同密度

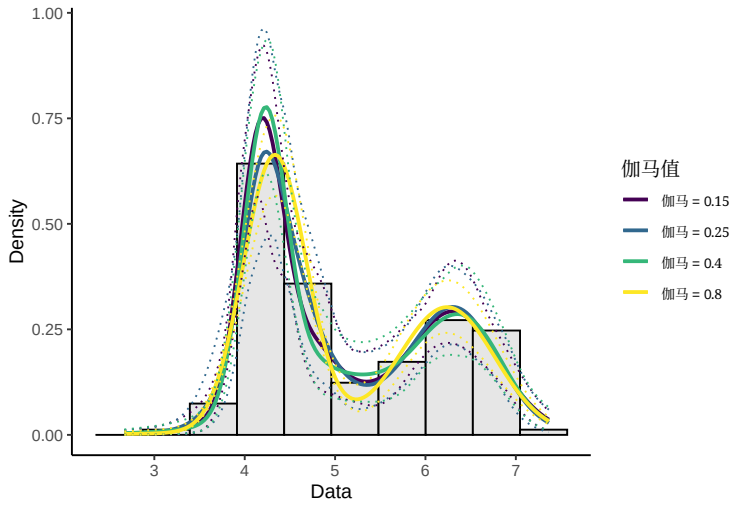

图 22： 模型对参数 γ的敏感性研究。此处采用了一个玩具数据集（来自 BNPdensity 的 acidity）。选择 γ 的取值是为了获得不同的先验行为。

采用不同 γ 值的模型非常相似。这一结果支持将 γ设为固定值的选择。本文中我们考虑 $\gamma = 0.4$。

C.2 参数 σ

我们对参数的先验分布选择σ也进行了类似的分析。可以考虑多种先验分布，尤其是均匀分布和截断正态分布。论文中呈现的结果是基于在集合 $[0, 1, 1, 5]$ 上采用均匀先验分布得到的。我们通过改变该先验设定的端点，研究了模型对这一设定的敏感性；此外，还考察了均值为 0。5、方差为 1、下界为 0。1 的左截断正态先验分布。结果见图 23。对于均匀先验分布，这里呈现的结果

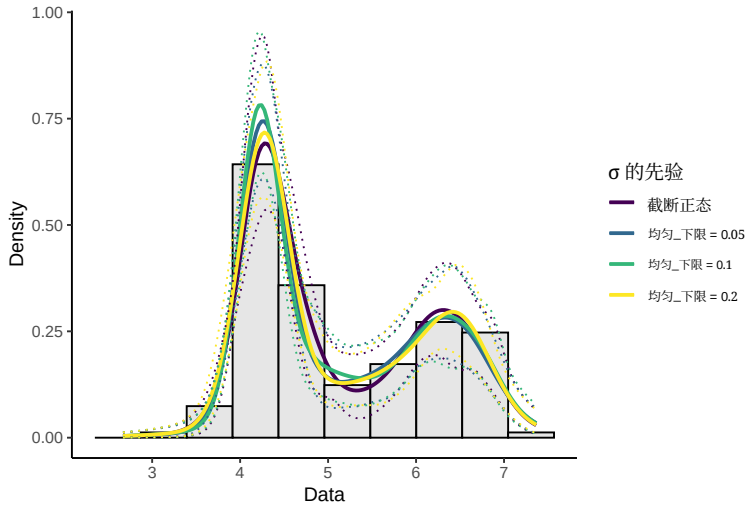

图 23： 研究模型对参数 σ先验选择的敏感性。此处使用了一个玩具数据集（来自 BNPdensity 的 acidity）。研究了 σ 的不同先验，例如具有不同边界的均匀分布和截断正态分布。

仅涉及修改下界的情况，但我们也考虑了上界，并得到了类似结果。根据这些结果，我们得出结论：模型对 σ 上先验分布的选择具有稳健性。需要指出的是，我们还可以考虑其他先验分布，但我们希望 σ 具备第2节所述的某些性质，这也解释了为何只考虑均匀分布和截断正态分布。

D 非负张量分解的详细信息

本文中，张量是矩阵的高维形式。我们仅考虑三阶张量，即具有三个维度的任意张量：$\mathbf{Y} \in \mathbb{R}^{I\times J\times K}$ 其中 I，J，$K \in \mathbb{N}$。我们介绍一种用于张量分解或分解的有用乘积，即外积。两个向量 $a \in \mathbb{R}^I$和 $b \in \mathbb{R}^J$的外积记为 $\circ$，得到一个矩阵 $A \in \mathbb{R}^{I\times J}$，

$$A = a \circ b = ab^T.$$

三个向量 $a \in \mathbb{R}^I$、$b \in \mathbb{R}^J$ 和 $c \in \mathbb{R}^K$ 的外积得到一个三阶张量$\mathbf{Y} \in \mathbb{R}^{I\times J\times K}$，

$$\mathbf{Y} = a \circ b \circ c, \qquad \text{with } y_{ijk} = a_i b_j c_k.$$

定义为三个向量外积的三阶张量称为秩一张量。张量$\mathbf{Y}$ 的秩定义为秩一张量的最小数量 R，即存在秩一张量$\mathbf{Y}_1$、.....、$\mathbf{Y}_R$，使得$\mathbf{Y} = \sum_{r=1}^{R} \mathbf{Y}_r$。

张量分解是矩阵分解技术的推广。不同的矩阵分解方法尤其适用于特征选择或降维。张量或多路数组分解能够用于处理数据包含高阶结构的应用场景。PARAFAC 模型是最常用的高阶张量分解模型之一。下面我们将介绍三阶张量的分解，但该模型也可以扩展到更高阶张量的分解。

D.1 PARAFAC 分解

我们回顾第 5 节中介绍的平行因子分析（PARAFAC）。给定张量$\mathbf{Y} \in \mathbb{R}^{I\times J\times K}$，PARAFAC 分解记作 $\mathbf{Y} = [\![A, B, C]\!]$ 其中 $A = [a_1, \ldots, a_R] \in \mathbb{R}^{I\times R}$，$B = [b_1, \ldots, b_R] \in \mathbb{R}^{J\times R}$和 $C = [c_1, \ldots, c_R] \in \mathbb{R}^{K\times R}$ 是三个分量矩阵。更正式地说，

$$\mathbf{Y} = \sum_{r=1}^{R} a_j \circ b_j \circ c_j + \mathbf{E} = [\![A, B, C]\!] + \mathbf{E},$$

其中，张量 $\mathbf{E} \in \mathbb{R}^{I\times J\times K}$ 表示近似误差。
PARAFAC 模型面临的一个难题是选择适当数量的组分R。这个问题等价于确定张量秩，因为张量秩是精确 PARAFAC 分解中 R 组分的最小数量，也就是说，近似误差张量 $\mathbf{E}$为零。已知确定给定张量的秩是一个 NP 难问题。

在理想情况下，数据中没有噪声，因此我们可以针对不同的 R 值拟合 PARAFAC 模型，直到获得精确分解。这假设存在完美的 PARAFAC 模型拟合程序，但实践中并非如此。此外，在更现实的情况下，数据包含噪声，因此所述程序不再适用。

已有多种方法可用于解决这一问题，例如核心一致性诊断、残差分析、观察载荷的视觉表现（也称为由分量矩阵表示的因子），或交叉验证。本文采用的方法是第5节所述的交叉验证。

D.2 非负张量分解

对于前文介绍的 PARAFAC 模型，可以施加一些非负约束。非负约束有助于赋予所发现的不同组分以物理意义。

实际上，在 PARAFAC 模型中添加约束，意味着需要求解一个带约束的优化问题。具体而言，求取 PARAFAC 分解意味着求解以下优化问题，

$$\min_{A,B,C} \|\mathbf{Y} - [\![A, B, C]\!]\|_F^2 ,$$

其中 $\|\cdot\|_F$表示由 $\|\mathbf{Y}\|_F^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} y_{ijk}^2$定义的张量 Frobenius 范数。非负 PARAFAC 分解是满足以下非负约束的 PARAFAC 分解

$$\min_{A,B,C \text{ s.t. } a_{ir}, b_{jr}, c_{kr}\in\mathbb{R}_+} \|\mathbf{Y} - [\![A, B, C]\!]\|_F^2 .$$

E 第 5 节的补充图5

现在，我们展示一些补充图，以说明第 5节中呈现的结果。
图 24 展示了使用交叉验证方法选择秩分解的过程。图 25 和图 26 展示了对张量分解所得的每个组分进行的K均值聚类。在第5节中考虑的7个组分，即分量A、B、C、D、E、F和G，是使用K均值聚类选出的，也是唯一在

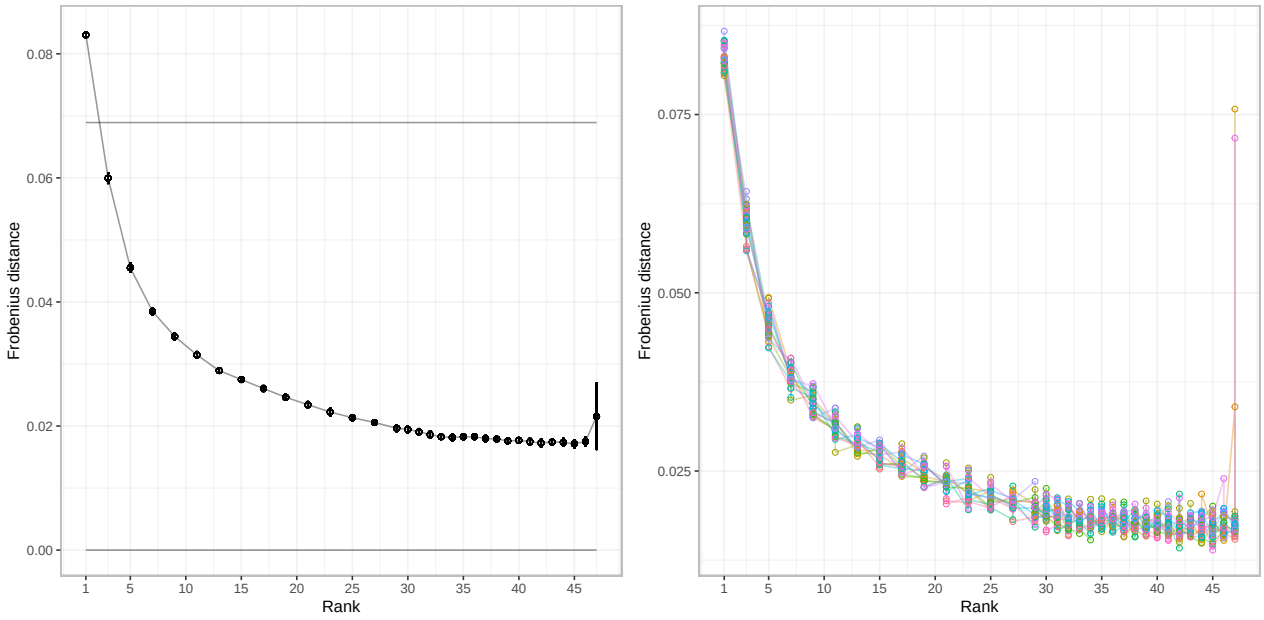

图24： 张量分解的交叉验证结果。左：秩在1至47之间变化时，30折交叉验证的弗罗贝尼乌斯距离均值，以及每个点周围的95%置信区间。右：30折交叉验证中每一折的弗罗贝尼乌斯距离。分解的秩通过以下方式选定：取误差最小对应秩（此处为45）的置信区间内包含的最低秩（此处为39）。

同时考虑污染物和物种的K均值聚类中良好分离的组分。这些选定的组分是分解中的第2、4、12、14、20、30和34个组分。
图 27 展示了组分与物种的热图，图 28展示了组分与污染物的热图。这些热图显示了哪些污染物或物种在某个组分中的权重较大。这是展示组分组成的另一种方式。

图 30 和 29 展示了此前选定的七个组分在两个不同分类等级上的分类组成。这两幅图支持这样一种观点：分类学似乎并不能显著决定物种敏感性。

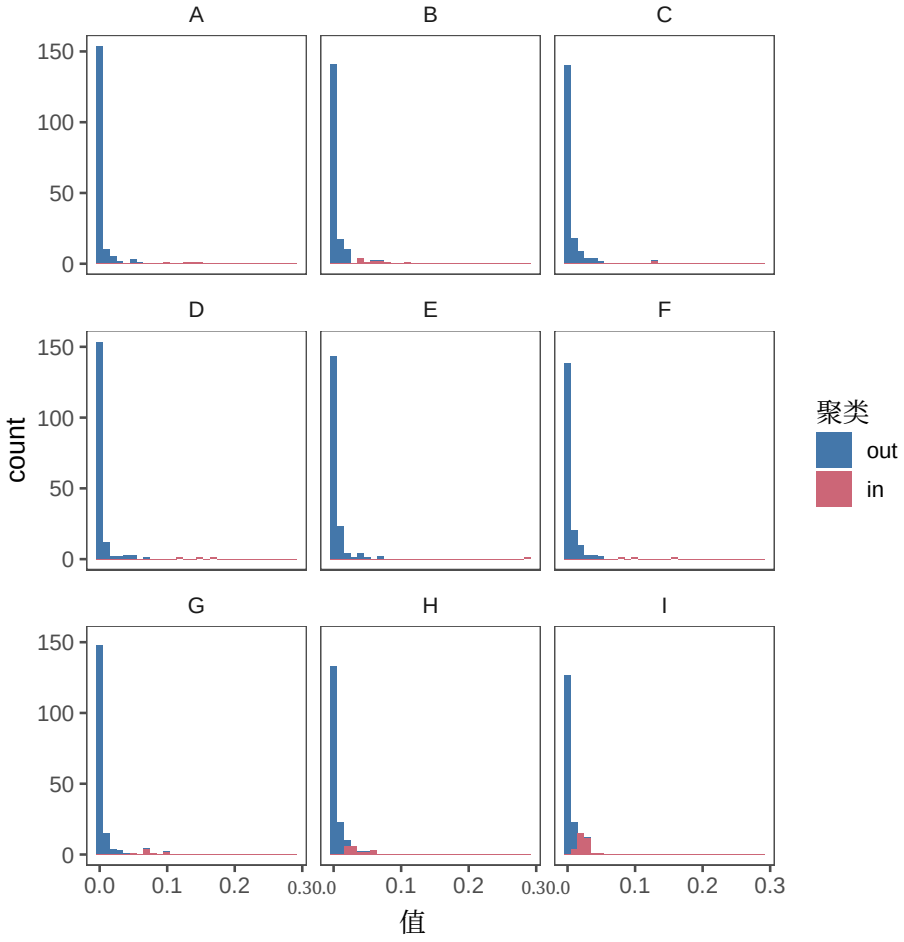

图 25：各组分中污染物权重的分布。分布中存在两个明显分离的组，这表明可以将属于该组分的污染物与不属于该组分的污染物区分开来。聚类采用 K 均值方法进行。在第 5 节中，仅考虑红色部分与蓝色部分没有重叠的组分，即组分 A、B、C、D、E、F 和 G，而不考虑组分 H 和 I。

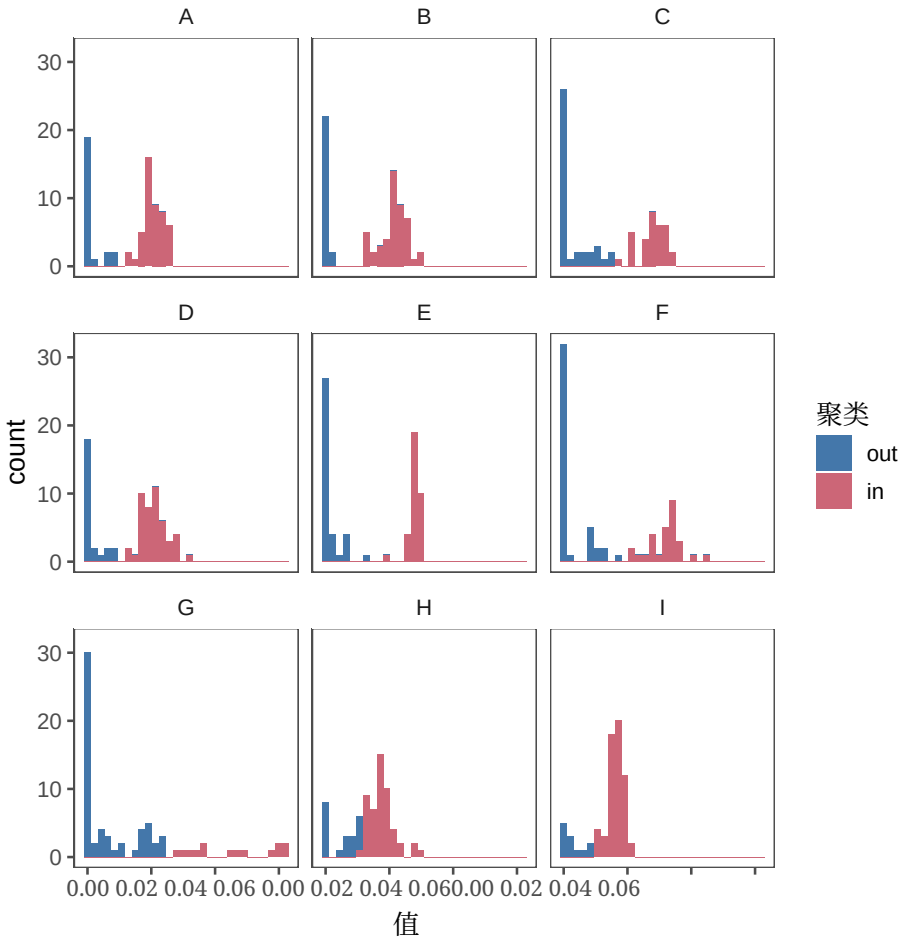

图26：各组分中物种权重的分布。该分布中存在两个明显分离的组，表明应将属于该组分的污染物与不属于该组分的污染物区分开来。聚类采用 K 均值方法进行。

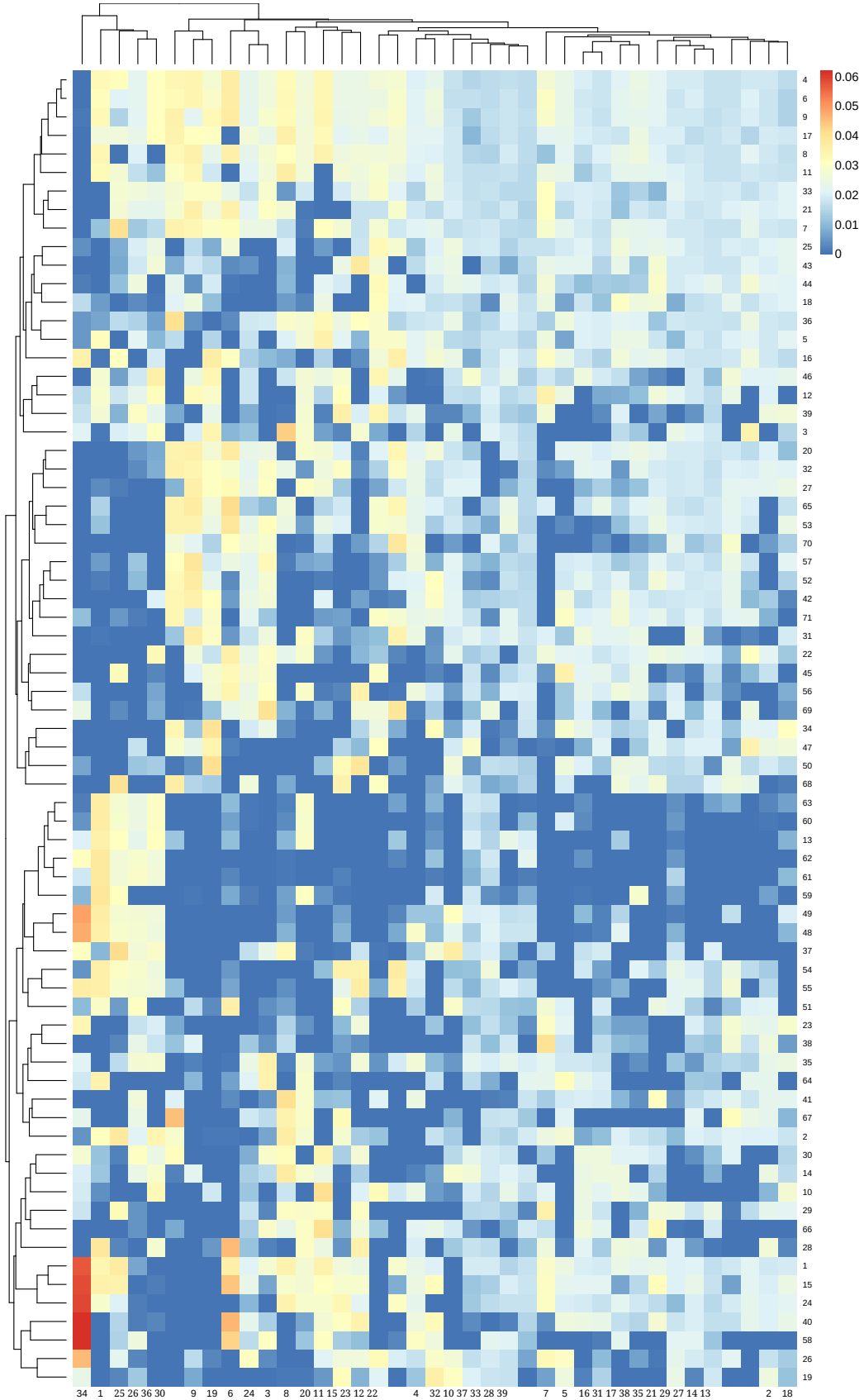

图 27: 所有物种和组分的热图。

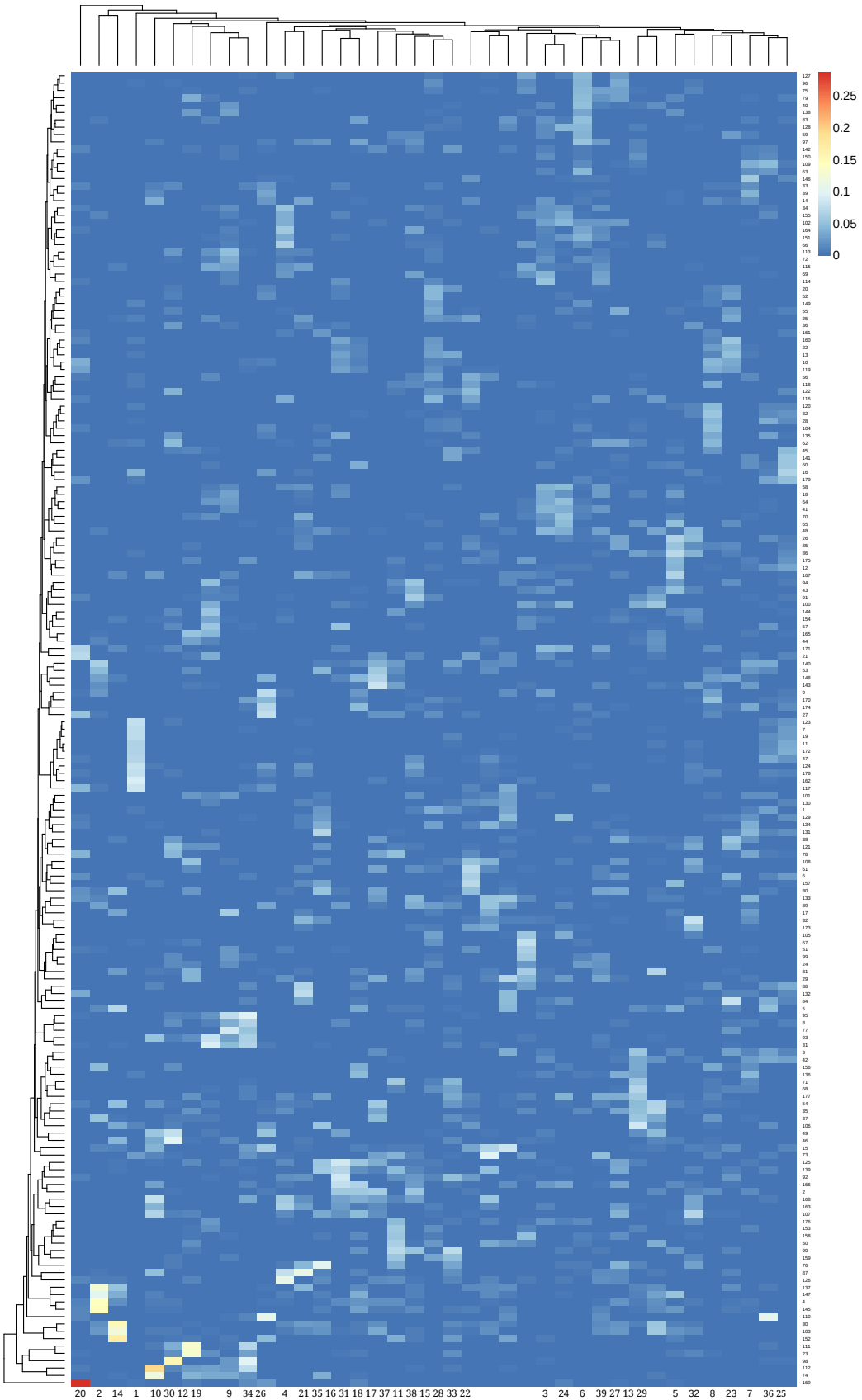

图28：所有污染物和组分的热图。

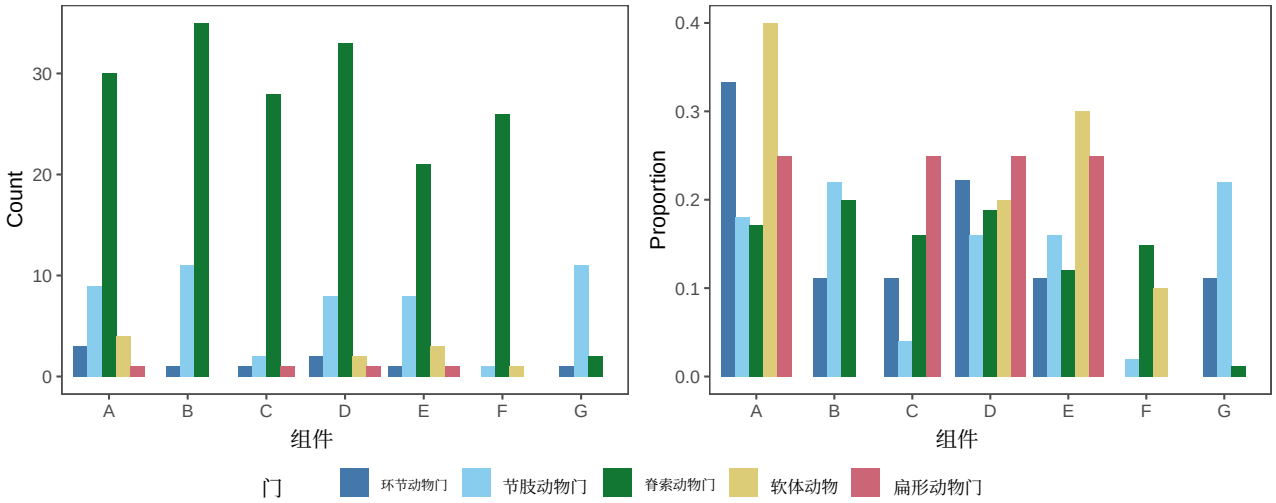

图29： 门水平上选定组分的分类组成。

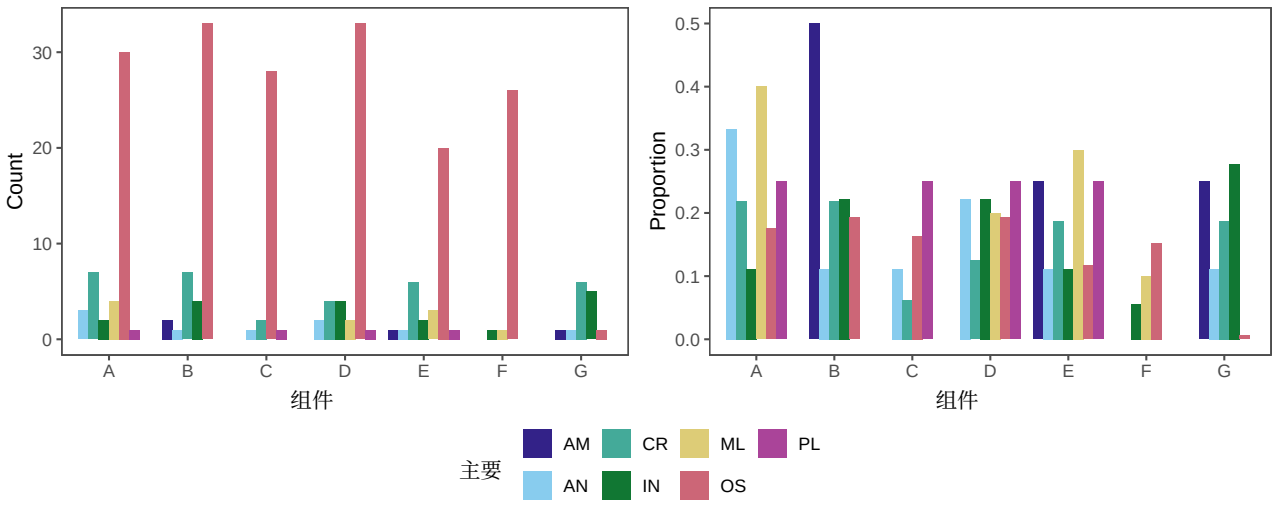

图30： 主要水平上选定组分的分类组成。