KernelMix：Python 中的贝叶斯核机器回归

Nilson Chapagain，得克萨斯农工大学统计学系

arXiv:2611.28634v1 [stat.CO] 21 Nov 2026

摘要

共同测量的暴露因素可能与结局存在非线性关联，而且某一暴露因素与结局的关联可能取决于另一暴露因素。贝叶斯核机器回归通过灵活的暴露–反应函数来表示这些模式。我们在 KernelMix 这一 Python软件包中实现了该方法。为处理不同类型的结局，KernelMix 将暴露函数与连续型、二元型、计数型和生存模型相结合，并纳入针对聚类观测和删失事件时间的特定扩展。对于拟合后的模型，该软件包可返回后验汇总、不确定性区间、联合预测和采样诊断。模型拟合与预测使用 NumPy 和 SciPy，无需依赖 R。为评估实现效果，我们将选定的高斯和 probit 后验汇总及条件计算结果与已安装的 R bkmr 0.2.2 进行比较，并通过独立的数学计算检验其他模型组件。每日臭氧分析展示了如何将这些组件整合为一个分析工作流程。

1 动机

暴露混合分析探讨多个共同测量的暴露因素如何影响结局。仅包含各暴露因素独立线性项的模型，无法表示曲线关联，也无法表示随另一暴露因素而变化的关联。贝叶斯核机器回归（BKMR）用联合暴露–反应函数取代这些项，因此无需预先指定每个非线性项和交互作用，就能表示这些模式[1]。

我们开发了 KernelMix，以便在 Python 中提供这一分析工作流程。该软件包将模型拟合与暴露对比、联合预测、不确定性汇总和采样检查相结合。NumPy [2]和 SciPy[3]负责这些计算，因此拟合和预测无需使用 R。我们的贡献在于软件实现及其经过文档说明的数值检查；贝叶斯核机器回归的建模方法由 Bobb 及其同事提出。

2 相关软件

R软件包bkmr提供高斯和 probit BKMR模型，以及暴露关联汇总 [4]。我们使用已安装的 bkmr 0.2.2 作为这些模型中部分计算的数值参照。我们通过构造参数状态来检查条件计算。我们还在数据和先验分布匹配的情况下比较拟合所得的后验汇总，并考虑有限采样带来的蒙特卡罗误差。

KernelMix 还支持计数和生存结局、其他聚类和残差模型，以及其他核表示。这些选项采用不同的似然、先验分布和预测规则，因此与高斯和 probit 参照模型结果一致，并不能验证这些选项。

3 方法与实现

我们首先介绍联合暴露函数如何纳入连续型结局模型。根据 Bobb 等人， [1], 令 Y_i 为观测 i的结局，$\mathbf{z}_i$ 为其 M 暴露因素，$\mathbf{x}_i$ 为其调整协变量。我们将联合暴露贡献写作 $h(\mathbf{z}_i)$，并通过系数 $\boldsymbol{\beta}$进行线性调整。这些项共同构成高斯模型

$$Y_i = h(\mathbf{z}_i) + \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta} + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2).$$

函数 h 包含 非线性关联和交互作用，而 ϵ_i 表示独立误差，其满足方差 σ^2。

拟合 h 需要一种方法来关联其在不同暴露组合下的取值。我们通过高斯过程（GP）先验来指定这种关系。先验描述的是使用结局数据之前的参数或函数取值；后验则利用数据更新该分布。对于 n 个观测值，写作 $\mathbf{h} = (h(\mathbf{z}_1)\dots, h(\mathbf{z}_n))^\mathsf{T}$。高斯过程先验通过高斯核 [1, 5],关联这些取值

$$\mathbf{h} \sim N(\mathbf{0}, \tau \mathbf{K}_{Z,\mathbf{r}}), \qquad K(\mathbf{z}, \mathbf{z}'; \mathbf{r}) = \exp\left\{ -\sum_{m=1}^{M} r_m (z_m - z'_m)^2 \right\}.$$

核函数 K 描述暴露组合之间的相似性，而 τ 控制 h的变化。每个 $r_m \geq 0$控制暴露 m 如何影响这种相似性。将 $r_m = 0$ 设为零，即可从 h中移除暴露 m，从而实现暴露选择。组选择会限制每个纳入组最多包含一种暴露 [1]。纳入概率概括了拟合模型下的选择情况；它们和 r_m 都不能衡量因果重要性。

上述高斯模型将暴露函数和线性调整与连续型结局结合起来。要处理其他结局，我们保留这两项贡献，但改变它们与结局分布之间的连接函数。对于二元结局，当一个潜在正态变量大于零时，Probit模型给出结局为 1 [4]。泊松模型和负二项模型用于描述计数数据 [6]。对于事件时间，分段指数模型采用在每个时间区间内保持恒定的基线风险 [7]。风险率是仍处于风险中的个体发生事件的瞬时速率 [6]。

事件时间模型的拟合还取决于每个事件的观测信息。生存记录可以注明个体在零时点之后入组、竞争事件原因，或事件仅知发生于某个时间区间内。提供的协变量历史在每个已记录的风险区间内保持恒定。对于延迟入组，推断针对的是纳入队列，无法还原入组前未观测风险的分布。

除了结局类型之外，重复观测值在同一聚类内也可能存在共同变异。随机截距和斜率通过共享系数 [4, 8] 来表示这种变异。对于高斯结局，残差模型可以分别指定不等的误差方差，或指定聚类内的相关性 [8]。

这些结局和依赖结构选项不同于用于关联暴露组合的核。高斯核可以替换为 Matérn $3/2$ 或有理二次核 [5]。为了减少 GP 坐标的数量，我们还提供预测过程和成对随机傅里叶表示 [9, 10]。这种降维会改变模型近似：预测过程会忽略在给定结点后剩余的变异，而随机傅里叶推断则以采样得到的特征为条件。大型残差相关区块和联合预测请求仍可能需要使用稠密矩阵。

这些模型选项也会影响采样和先验约定。默认的高斯采样器会在参数采样期间对 $\mathbf{h}$ 进行解析积分；默认的高斯拟合和 Probit 拟合会存储 $\lambda = \tau/\sigma^2$，其中 Probit 拟合使用 $\sigma^2 = 1$。计数、生存和结构化拟合则存储直接振幅 τ 作为 lambda，并对调整系数使用正规正态先验。这些拟合中的高斯先验区块采用椭圆切片采样[11]。因此，比较先验时需要依据其标明的参数尺度。手册说明了各选项的输入、先验和预测规则。

4 软件检查

我们在四个固定数据集上，将拟合后的验后汇总结果与已安装的 R bkmr 0.2.2 进行比较：高斯模型和 Probit模型各一个，且每种模型分别包含和不包含随机截距。两个实现使用完全相同的观测值、暴露标准化、调整设计和先验分布，并禁用暴露选择。每个实现运行四条链，提供参数、 h的三个取值以及一个暴露对比量的均值和第 10、50、90 后验百分位数。每条 R 链在 3,000 次预热迭代后保留 120,000 个抽样值；每条 Python 链保留 60,000 个。所有 160 个检查汇总结果均通过固定的诊断、精度和差异规则。我们要求秩-$\hat{}$归一化/折叠分割 $R \leq 1.01$、整体/尾部有效样本量至少为 400，且蒙特卡罗标准误不超过合并后验标准差的 2%。手册给出了完整的比较规则和结果；这些数值筛查并不能证明后验等价性或一般意义上的收敛。此外，构造状态检查还会比较条件均值、协方差、纳入概率和先验密度。

对于没有 R 参考实现的模型，我们通过独立数值积分比较选定的后验矩。三个固定案例涵盖负二项计数、右删失生存数据，以及包含随机截距/斜率和 AR(1)残差相关的高斯模型。所有案例均使用常数暴露函数，并对其高斯过程振幅进行抽样。结构化案例固定随机系数协方差和残差相关性。全部 12 个

个已检查的矩和四个选定的协方差均满足误差和精度规则。这些小型案例用于评估所述后验量，而非每一种结构化选项或一般性的收敛情况。

已安装的 wheel 在两个原生 macOS 环境中均通过了 512 项测试：Python 3.14 搭配当前依赖项，以及 Python 3.10 搭配最低依赖项。已配置的托管 Linux 和 Windows 任务在运行器启动前便已停止，因此 Windows 和托管 Linux 的兼容性仍未得到验证。

数值一致性不同于复现原始的贝叶斯核机器回归研究。原始模拟在每种情景下使用了300个数据集，$n = 100$ 并对经验暴露行进行重采样；应用部分分析了375名儿童，以及来自13只狗的142条暴露记录 [1]。在所检查的开放资源中，未能找到原始的孟加拉国、波士顿和狗数据集，也未能找到足够的拟合设置。具有相似协方差的高斯生成的暴露将构成不同的实验。因此，我们使用下面的观测数据示例来展示工作流程，但不将其视为原始研究复现或区间覆盖率研究。

5 观测数据示例

5.1 数据

我们使用公共空气质量数据，考察在针对日历月份进行调整后，每日臭氧如何随太阳辐射、风和温度变化 [12, 13]。这些数据包含1973年5月至9月纽约的153条每日记录。臭氧和气象测量值来自不同的站点和时间，因此并不描述同一个人的联合暴露。

在拟合模型之前，我们确定了所有必需测量值均完整的天数：保留111天，排除42天。其中有37个缺失的臭氧值和7个缺失的太阳辐射值，且两者存在重叠。图1显示，这种排除在各月份之间并不均衡：六月仅保留30天中的9天，而九月保留30天中的29天。对于这些保留的天数，图2展示了臭氧值与各项天气测量值的关系。这些成对图未对其他气象变量和月份进行调整。

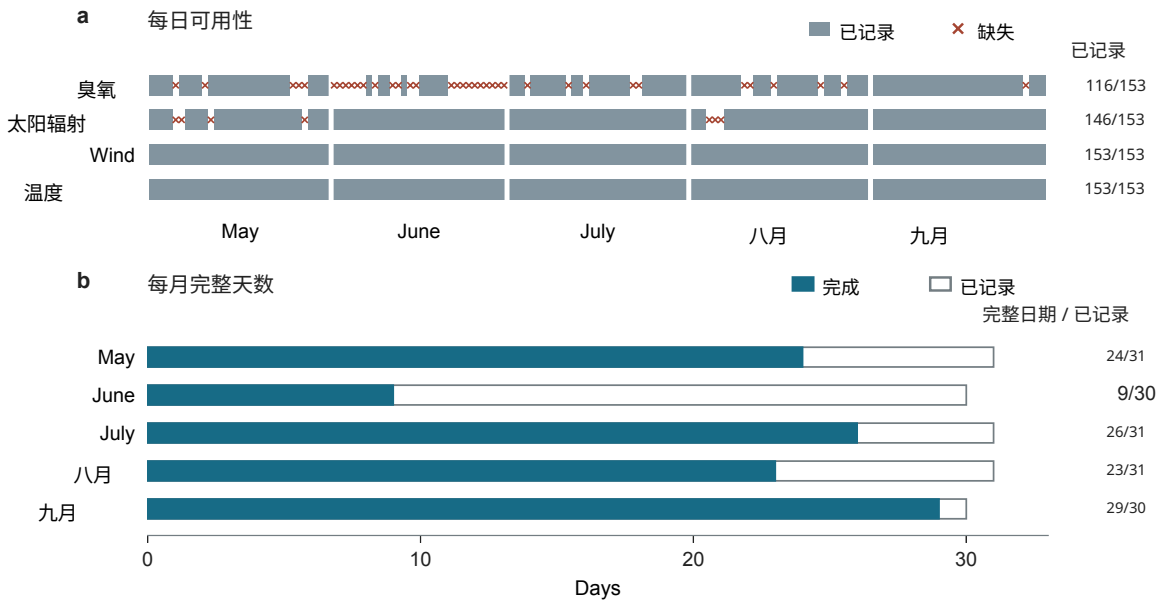

图1：1973年5月至9月153条每日记录中的测量可用性。（a）每个单元格代表一天：填充单元格表示已记录测量值，叉号表示缺失值。分数表示已记录测量值占153天的比例。（b）填充条形表示完整日期，轮廓表示所有已记录日期；分数表示完整日期计数/已记录日期计数。分析使用111个完整日期。缺失值既未进行插补，也未按零处理。

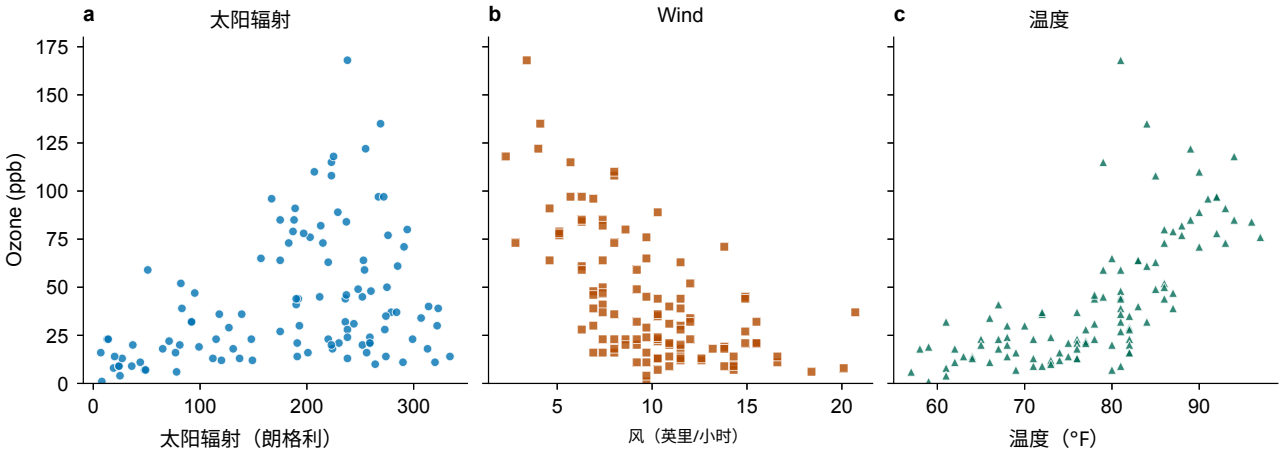

图2：111个完整日期中，臭氧与太阳辐射、风和温度的关系。每个点代表一天。图中显示的是未经调整的观测值，没有控制其他气象变量或月份。

5.2 模型与拟合关联

为了同时估计天气关联并对月份进行调整，我们对 $Y_i = \log(\text{Ozone}_i)$ 拟合高斯模型，其中太阳辐射、风和温度位于 $\mathbf{z}_i$ 中。调整向量 $\mathbf{x}_i$ 包含截距项以及六月至九月的四个指示变量，以五月为参照。我们使用保留天数对暴露因素进行标准化。这种月份调整反映了广泛的季节性差异；独立误差假设未考虑连续天数之间的依赖性。

我们使用相互独立的 $r_m \sim \text{伽马}(1,\ 1)$ 先验、用于 σ^2 的逆伽马(2，1) 先验、 $\lambda \sim \text{伽马}(1,\ 0。1)$ 先验，以及用于 β的默认平坦先验。伽马参数分别为形状参数和速率参数。暴露选择已禁用，因此本分析估计联合暴露函数，而不选择气象变量的子集。

在指定模型和先验分布后，我们使用随机种子 73 依次运行了四条链。每条链均先进行 5,000 次丢弃的预热迭代，随后保存 20,000 个抽样值。一次初始运行保存了 4,000 个抽样值，其最大 $\widehat{R} = 1.0142$；之后我们在不更改模型、数据或先验分布的情况下延长了采样。曲线和拟合残差汇总使用每条链中均匀间隔抽取的 400 个抽样值。拟合在已记录的本地环境中耗时 72.6 秒。

在较长的运行中，最大 $\widehat{R}$ 为 1.0053，最小主体有效样本量为 885，最小尾部有效样本量为 1,193。这些汇总涉及十个单独的参数；完整诊断结果如下所示。

图 3 展示了各项拟合天气关联：每次改变一种暴露，同时将其他暴露固定在其中位数。后验均值曲线通常随温度升高而上升，随风速增大而下降；太阳辐射曲线则呈非线性。这些是拟合的条件关联，并给出了独立误差模型下的逐点不确定性。

所展示的对比量为 $h(\mathbf{z}) - h(\mathbf{z}_{0.5})$，其中 $\mathbf{z}_{0.5}$ 是暴露中位数向量。在对数臭氧尺度上，对该对比量取指数可得两个暴露组合下拟合臭氧中位数之比，其中月份保持不变。它并不表示算术均值之差。

为考察拟合后剩余的变异，我们查看拟合值与观测值面板以及残差面板。这些数据使用了拟合时所用的同一批天数，因此评估的是拟合样本，而非对新数据的预测。在连续日历日的 92 对残差中，样本相关性为 0.0136；如此微弱的相关性并不能证明误差相互独立。

拟合模型将暴露函数与月份调整分开。图 4 分别报告两者的参数，区分核相关性 r_m 和月份系数 β。由于两者作用不同，因此需要使用不同的坐标轴。每个 95% 可信区间都包含拟合模型下 95% 的后验概率。由于未启用暴露选择，正的 r_m 值描述的是拟合核，并不能作为变量选择证据。

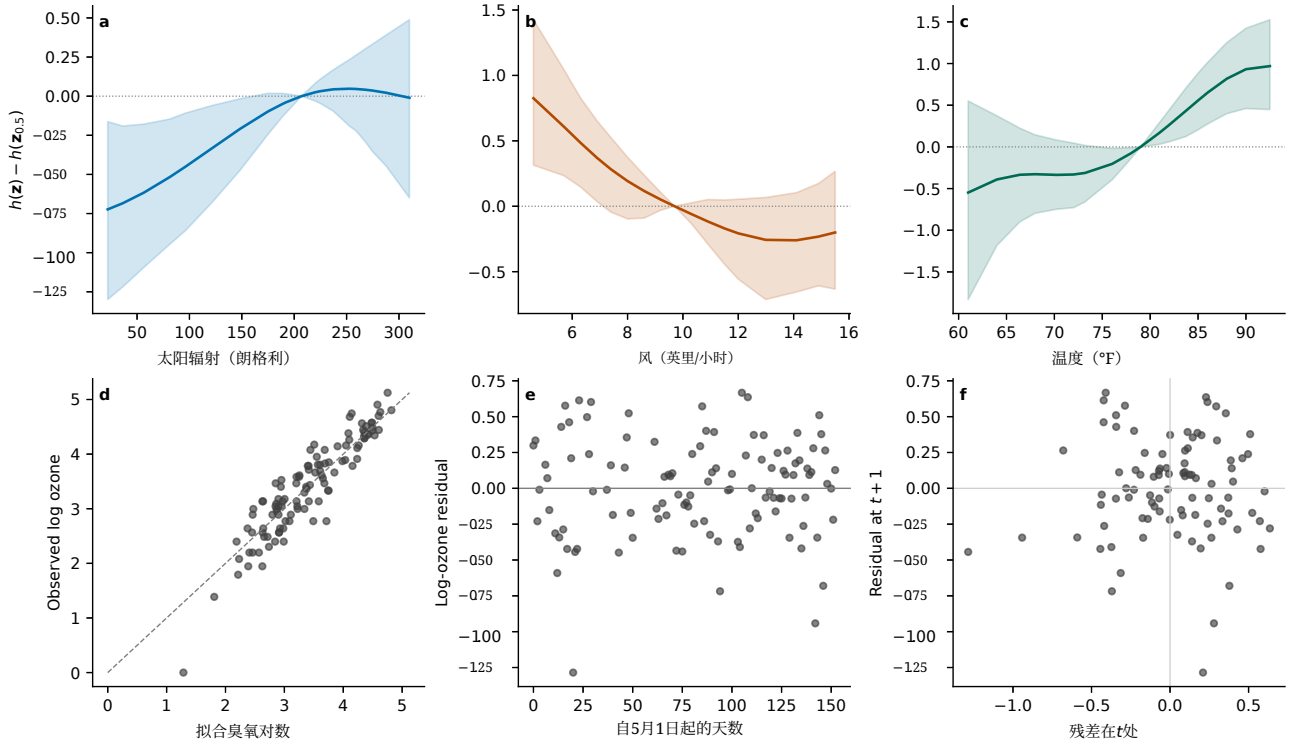

图 3：拟合的暴露关联与残差（$n = 111$）。（a–c） $h(\mathbf{z}) - h(\mathbf{z}_{0.5})$ 的后验均值和 95% 逐点可信区间，其中其他暴露因素固定在各自的中位数，并且 $\mathbf{z}_{0.5}$ 全中位数剖面。（d–f）拟合与观测对数臭氧、按日期排列的残差，以及相隔一个日历日的残差对。缺失日期之间不连线。区间假设工作误差相互独立且服从高斯分布。

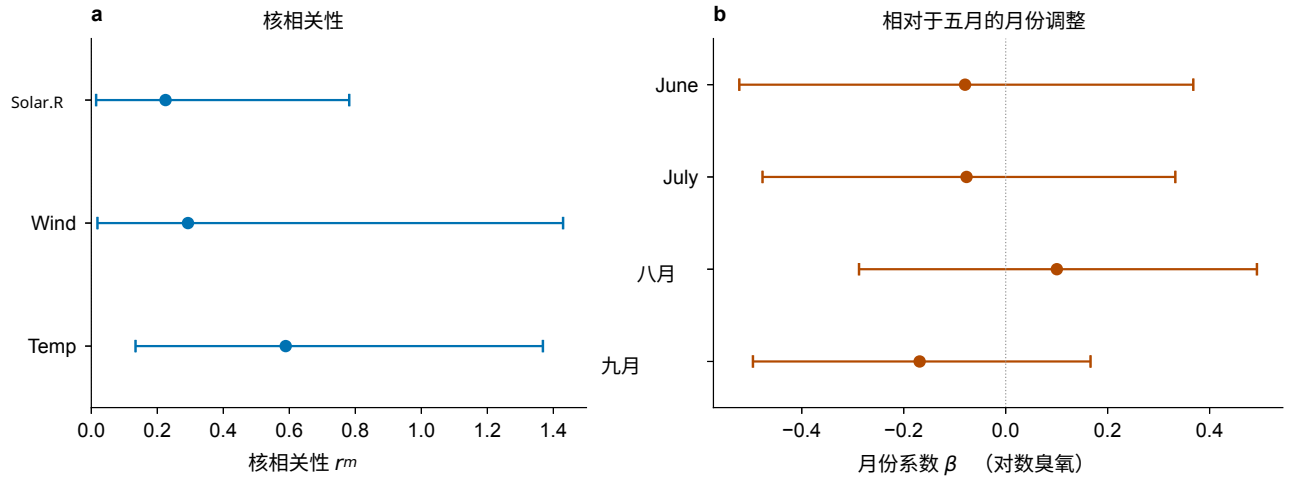

图 4： β中（a）核相关性参数 r_m 和（b）月份系数相对于五月、在对数臭氧尺度上的后验均值和 95% 等尾可信区间。截距未显示。各面板的刻度不同；核相关性并不衡量因果重要性。

5.3 采样诊断与暴露支持度

前面的区间取决于后验采样是否充分。我们使用分裂 $\hat{R}$进行采样诊断，该方法比较各链；同时考察有效样本量，以考虑抽样值之间的依赖关系 [14] （图 5）。主体有效样本量反映后验分布的主体部分，而尾部有效样本量反映其尾部。这些检查涵盖十个单独参数；即使结果良好，也不能证明每条暴露曲线均已收敛，或模型假设有效。

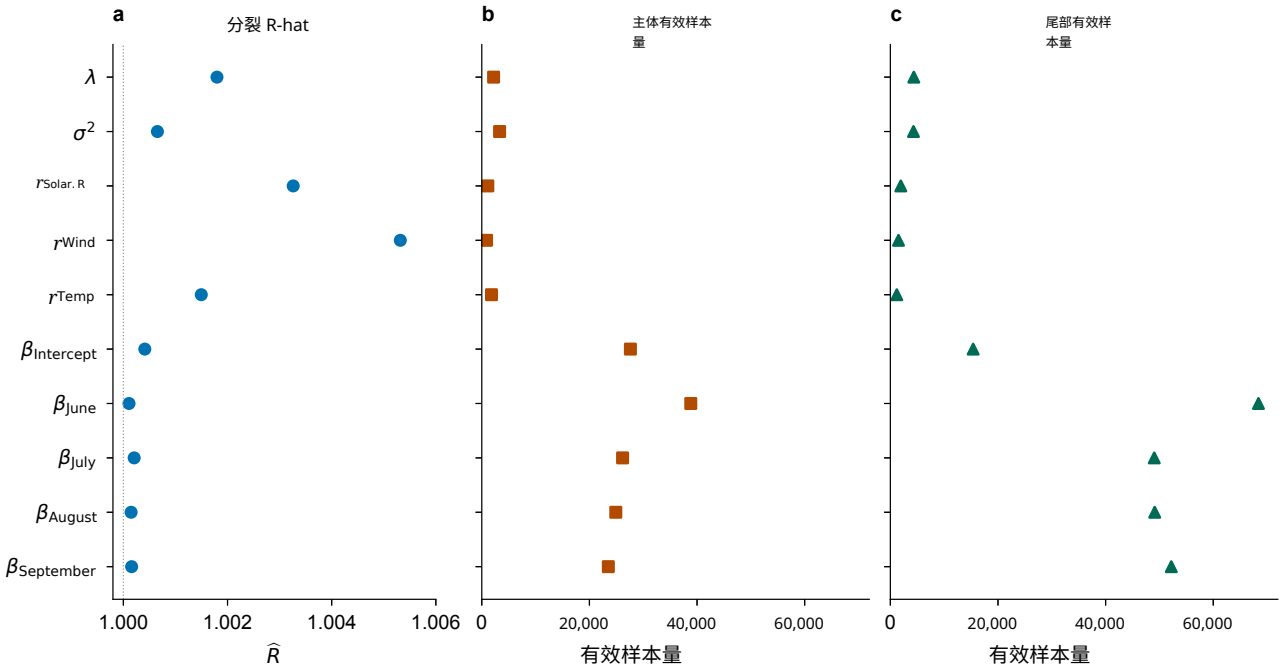

图 5：对四条链中十个参数的采样诊断，每条链均保存了 20,000 个抽样值。(a) 秩-$\hat{R}$标准化/折叠分裂 $\hat{R}$。(b) 主体有效样本量。(c) 尾部有效样本量。这些诊断评估的是采样，而非模型假设。

充分的采样也不能确保绘制的暴露组合在数据中实际出现。每条曲线使用该暴露因素的第 5–95 百分位数，并去除重复值，同时将其他暴露因素固定在其中位数。这种构造可能产生与观测行相距甚远的组合。因此，我们报告经过训练标准化后的最近的完整暴露组合距离（图6）。距离越大，表示附近数据越少；我们不使用任何阈值来判定预测是否可靠。

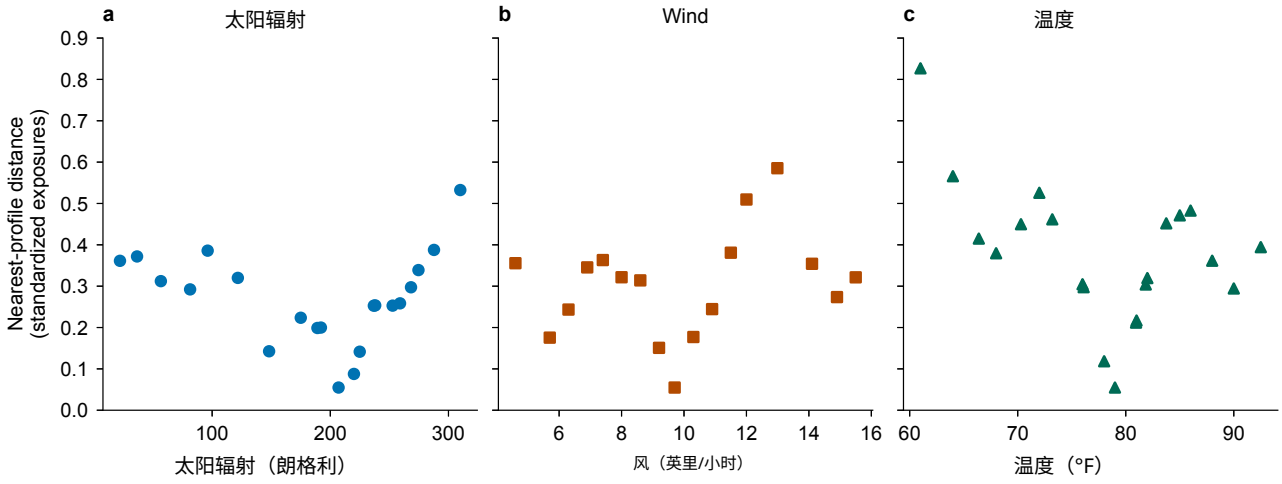

图6：暴露标准化后的最近观测组合距离。每个绘制值都与其他暴露因素的中位数结合。网格包含21个太阳辐射值、17个风值和21个温度值。该距离衡量其与完整观测数据的接近程度，不设模型有效性阈值。

6 限制与可复现性

臭氧曲线描述了在指定先验和独立高斯误差下的条件关联。月份调整并未对日间依赖性进行建模，而完整病例分析可能引入选择偏倚[15]。采样诊断无法解决这两个问题，因此该分析不支持任何因果或预测主张。数值检查也未能解决重复数据集偏倚、区间覆盖率、原始研究复现和大数据性能问题。

为使这些计算和臭氧工作流程具有可复现性，位于 `https://github.com/nilson01/kernelmix` 的代码仓库包含安装说明、函数参考、测试、源数据和分析脚本。保存的表格记录了数据哈希值、拟合设置、环境和诊断结果。图形脚本使用 Matplotlib [16]。该软件依据 GPL-2.0 发布；R 参考资料和数据来源另行注明。

致谢

计算工作在德州农工大学统计学系的 Arseven 计算集群上完成。

本研究未获得任何资助。

参考文献

[1] Bobb JF 等（2015）。利用贝叶斯核机器回归估计多污染物混合物的健康影响。《生物统计学》16，493–508。doi:10.1093/biostatistics/kxu058。[2] Harris CR 等（2020）。使用 NumPy 进行数组编程。《自然》585，357–362。doi:10.1038/s41586-020-2649-2。[3] Virtanen P 等（2020）。SciPy 1.0：Python 科学计算的基础算法。《自然·方法》17，261–272。doi:10.1038/s41592-019-0686-2。[4] Bobb JF、Claus Henn B、Valeri L 和 Coull BA（2018）。用于通过贝叶斯核机器回归分析多重并发暴露健康影响的统计软件。《环境健康》17，67。doi:10.1186/s12940-018- 0413-y。[5] Rasmussen CE 和 Williams CKI（2006）。《机器学习中的高斯过程》。麻省理工学院出版社。第4章。[6] Stan 开发团队。Stan 文档，2.40 版。计数分布；生存模型。[7] Martino S 和 Rue H（2009）。使用集成嵌套拉普拉斯近似实现近似贝叶斯推断：inla 程序手册，第3.15节。手册。

[8] Pinheiro JC 和 Bates DM。nlme 文档，3.1-168 版。混合效应模型；AR(1)相关；指数方差。[9] Banerjee S、Gelfand AE、Finley AO 和 Sang H（2008）。适用于大型空间数据集的高斯预测过程模型。《皇家统计学会志：B辑》70，825–848。doi:10.1111/j.1467-9868.2008.00663.x。[10] Rahimi A 和 Recht B（2007）。面向大规模核机器的随机特征。《神经信息处理系统进展》第20卷。会议论文。[11] Murray I、Adams RP 和 MacKay DJC（2010）。椭圆切片采样。《机器学习研究论文集》9，541–548。会议论文。[12] R核心团队（2025）。R：一种用于统计计算的语言与环境。R统计计算基金会，维也纳。数据导出：R 4.5.0，datasets::airquality。数据集文档。[13] Chambers JM、Cleveland WS、Kleiner B 和 Tukey PA（1983）。《数据分析的图形方法》。Wadsworth & Brooks/Cole。[14] Vehtari A、Gelman A、Simpson D、Carpenter B 和 Bürkner PC（2021）。秩正态化、折叠与局部化：一种改进的 $\widehat{R}$ 用于评估 MCMC 收敛的方法。《贝叶斯分析》16，667–718。doi:10.1214/20-BA1221。[15] Sterne JAC 等（2009）。流行病学和临床研究中缺失数据的多重插补：潜力与陷阱。BMJ 338，b2393。doi:10.1136/bmj.b2393。[16] Hunter JD（2007）。Matplotlib：一个二维图形环境。《科学与工程计算》9(3)，90–95。doi:10.1109/MCSE.2007.55。