真实世界分布偏移下，RegressionModel 的不确定性有多可靠?

FredrikK. Gustafsson瑞典乌普萨
拉大学信息技术系

Martin Danelljan瑞士
苏黎世联邦理工学院计算机
视觉实验室

ThomasB. Schön瑞典乌普萨拉大
学信息技术系

在 OpenReview 上评审： *https://openreview.net/forum?id=WJt2Pc3qtI*

arXiv:2302.03679v2 [cs.LG] 7 Nov 2023

摘要

许多重要的计算机视觉应用天然可表述为回归问题。在医学影像领域，精准的回归模型有望实现多种任务的自动化，从而降低成本并改善患者预后。然而，要在安全关键型部署中使用这些模型，就必须可靠地估计模型不确定性，同时还要应对实践中可能遇到的各种分布偏移。受此启发，我们着手研究回归不确定性估计方法在各种真实世界分布偏移下的可靠性。为此，我们提出了一个规模庞大的基准测试，涵盖 8 个基于图像的回归数据集，其中包含不同类型的挑战性分布偏移。随后，我们利用该基准测试评估了多种最常见的不确定性估计方法，以及两个源自分布外检测任务的最先进不确定性分数。我们发现，在没有分布偏移时，这些方法的校准效果良好；但在许多基准数据集上，它们都会变得过度自信。这一发现揭示了当前不确定性估计方法的重要局限，因此，拟议基准测试也为研究界提出了一项挑战。我们希望这一基准测试能推动更多研究，探索如何开发真正可靠的回归不确定性估计方法。代码见 `https://github.com/fregu856/regression_uncertainty`。

1 引言

回归是一个基础性的机器学习问题，在计算机视觉领域有许多重要应用（Rothe 等人，2016年；Law & Deng, 2018年；Xiao 等人，2018年；Zhu 等人，2018年；Shi 等人，2019年；Lathuilière 等人，2019年）。一般而言，它需要根据给定的输入 y 预测连续目标 x。在医学影像领域，许多任务都可以自然地表述为回归问题，包括脑龄估计（Cole 等人，2017年；Jónsson 等人，2019年；Shi 等人，2020年）、心血管体积和风险因素的预测（Xue 等人，2017年；Poplin 等人，2018年），以及身体成分分析（Langner 等人，2020年；2021年）。如果能够将机器学习模型部署到现实世界的临床实践中，自动回归此类多种属性，最终将有助于降低整个医疗系统的成本并改善患者结局（Topol，2019年）。

然而，在医疗应用和其他安全关键领域进行现实世界部署，对此类回归模型提出了极高的要求。尤其是，常见做法是训练一个深度神经网络（DNN），直接输出预测的回归目标 $\hat{y} = f(x)$，但这并不足够，因为它无法捕捉预测结果 $\hat{y}$ 中任何不确定性度量。因此，该模型无法检测例如输入 x

与其训练数据相比属于分布外（OOD）的输入。由于深度神经网络在分布外输入上的预测准确率通常会显著下降（Hendrycks & Dietterich, 2019; Koh 等，2021），这可能造成灾难性后果。因此，研究人员投入了大量精力，探索训练不确定性感知深度神经网络模型的各种方法（Blundell 等人，2015; Gal, 2016; Kendall & Gal, 2017; Lakshminarayanan 等人，2017; Romano等，2019），以明确估计预测结果中的不确定性。

然而，这些不确定性估计必须准确可靠。否则，如果模型偶尔变得过度自信，给出置信度很高却不正确的预测，那么提供不确定性估计可能只会带来虚假的安全感——甚至可以说，这会使模型更不适合安全关键型部署。具体而言，不确定性估计必须经过良好校准，并与预测误差恰当对应（Gneiting等，2007；Guo 等人，2017）。此外，即使在实际部署期间可能遇到的各种分布偏移下，不确定性也必须保持良好校准（Quionero-Candela 等人，2009；Finlayson 等人，2021）。例如，在医疗应用中，假设某个模型仅使用2020年从一家大型城市医院收集的数据进行训练，那么到了2023年，它也应该能为来自城市和农村地区的患者给出校准良好的预测。虽然研究人员已在分类任务中评估了分布偏移条件下的不确定性校准以及深度神经网络的一般鲁棒性（Hendrycks 等人，2021a；Ovadia 等，2019），但这一重要问题在回归领域尚未得到充分研究。

受此启发，我们着手研究回归不确定性估计方法在各种真实世界分布偏移下的可靠性。为此，我们提出了一个大型基准测试，其中包含 8 个具有不同类型分布偏移的基于图像的回归数据集（见图 1）。这些数据集均公开可用且规模相对较大（训练图像数量为 6 592–20 614），同时便于存储和训练模型（64×64 图像，回归目标为一维）。其中四个数据集还来自医疗应用，涉及临床相关分布偏移。我们评估了一些最常用的回归不确定性估计方法，包括保形预测、分位数回归，以及通常被视为最先进方法的集成（Ovadia 等，2019；Gustafsson 等人，2020b）。我们还考察了选择性预测方法（Geifman & El-Yaniv，2017）：在这种方法中，回归模型可以拒绝为某些输入输出预测。借助这一方法，我们能够评估分布外检测相关文献中的多种不确定性分数（Salehi 等人，2022）。具体而言，我们评估了近期提出的两种基于特征空间密度的分数（Mukhoti 等人，2021a；Pleiss 等人，2020；Sun 等，2022；Kuan 和 Mueller，2022）。

总的来说，我们评估了 10 种不同的方法。我们发现，其中没有任何一种方法在所有数据集上都接近完美校准。这些方法在没有分布偏移的基线变体上校准良好，但在我们的许多基准测试数据集上都变得过度自信。此外，尽管保形预测方法通常被宣传具有理论保证，但它们同样受到这一问题的影响。这凸显了始终了解潜在假设、评估这些假设在实践中是否可能成立的重要性。基于最先进的分布外不确定性分数的方法相对于其他方法表现良好，但在许多情况下也过度自信——其绝对性能可以说仍然不足。因此，我们拟议的基准为研究界提出了一项挑战，希望能够推动更多关于如何开发真正可靠的回归不确定性估计方法的研究。

贡献总结 我们收集了 8 个大规模且便于使用的基于图像的回归数据集，这些数据集包含不同类型的具有挑战性的分布偏移。基于此，我们提出了一个基准，用于测试回归不确定性估计方法在真实世界分布偏移下的可靠性。随后，我们利用该基准评估了许多最常见的不确定性估计方法，以及两种用于分布外检测的最先进不确定性分数。我们发现，所有方法在许多基准测试数据集上都变得过度自信，从而揭示了当前不确定性估计方法的局限性。

2 背景

在一个回归问题中，任务是针对任意给定的输入 $x^\star \in \mathcal{X}$，预测目标 $y^\star \in \mathcal{Y}$。为此，我们还获得了一个由独立同分布（i.i.d.）的输入-目标对组成的训练集，即 $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^N$、$(x_i, y_i) \sim p(x, y)$。回归与分类的区别在于，目标空间 $\mathcal{Y}$ 是连续的，即 $\mathcal{Y} = \mathbb{R}^K$。在本文中，我们只考虑一维情况，即 $\mathcal{Y} = \mathbb{R}$。此外，这里的输入空间 $\mathcal{X}$ 始终对应于图像空间。

预测区间、覆盖率与校准 给定期望的失覆盖率 α，预测区间$C_\alpha(x^\star) = [L_\alpha(x^\star)\ U_\alpha(x^\star)] \subseteq \mathbb{R}$ 是一个将输入 $x^\star$ 映射到某个区间的函数，该区间应以概率 $1-\alpha$覆盖真实的回归目标 $y^\star$。对于由 $N^\star$ 个样本构成的任意集合 $\{(x_i^\star, y_i^\star)\}_{i=1}^{N^\star}$，经验区间覆盖率 是指预测区间覆盖对应目标的输入所占的比例，

$$\text{Coverage}(C_\alpha) = \frac{1}{N^\star}\sum_{i=1}^{N^\star}\mathbb{I}\{y_i^\star \in C_\alpha(x_i^\star)\}. \tag{1}$$

如果覆盖率等于 $1-\alpha$，我们就称预测区间已完美校准。除非另有说明，本文中我们设定 $\alpha = 0.1$。因此，预测区间的覆盖率应达到 90%。

2.1 回归不确定性估计方法

基于图像的回归最常见的方法是训练一个深度神经网络 $f_\theta\colon \mathcal{X} \to \mathbb{R}$，使其能够针对任意输入 x 输出预测目标 $\hat{}y = f_\theta(x)$，例如使用 L2 或 L1 损失（Lathuilière 等人，2019年）。我们关注的是能够扩展这种标准直接回归方法、同时为预测提供不确定性估计的方法。具体而言，我们考虑针对每个输入 x 输出预测区间 $C_\alpha(x)$ 和预测目标 $\hat{}y(x) \in C_\alpha(x)$ 的方法。预测 $\hat{}y(x)$ 的不确定性由区间 $C_\alpha(x)$ 的长度量化（区间越长，不确定性越高）。一些最常用的回归不确定性估计方法属于这一类，下文将对此进行更详细的介绍。

保形预测 标准的回归方法可以通过采用拆分保形预测框架加以扩展（Papadopoulos 等人，2002；Papadopoulos，2008；Romano 等人，2019）。该方法将训练集$\{(x_i, y_i)\}_{i=1}^N$ 拆分为适当训练集$\mathcal{I}_1$ 和校准集$\mathcal{I}_2$。深度神经网络 f_θ在$\mathcal{I}_1$, 上进行训练，并在校准集$\mathcal{I}_2$上计算绝对残差 $R = \{|y_i - f_\theta(x_i)|\colon i \in \mathcal{I}_2\}$。给定一个新的输入 $x^\star$，随后根据预测值 $f_\theta(x^\star)$构造预测区间 $C_\alpha(x^\star)$，如下所示：

$$C_\alpha(x^\star) = [f_\theta(x^\star) - Q_{1-\alpha}(R, \mathcal{I}_2),\ f_\theta(x^\star) + Q_{1-\alpha}(R, \mathcal{I}_2)], \tag{2}$$

其中，$Q_{1-\alpha}(R,\mathcal{I}_2)$是绝对残差 R的第（$1-\alpha$）个分位数。假设训练数据和测试数据是可交换抽取的，则可以保证该预测区间满足 $\mathbb{P}\{y^\star \in C_\alpha(x^\star)\} \geq 1-\alpha$（边际覆盖保证）。对于所有输入 $x^\star$，区间 $C_\alpha(x^\star)$的长度均固定为 $2Q_{1-\alpha}(R,\mathcal{I}_2)$。

分位数回归深度神经网络还可以采用分位数回归方法进行训练，直接输出长度取决于输入的预测区间（Koenker & Bassett Jr，1978年；Romano 等人，2019年；Jensen 等人，2022年）。这需要估计条件分位数函数 $q^\alpha(x) = \inf\{y \in \mathbb{R}\colon F_{Y|X}(y|x) \geq \alpha\}$，其中 $F_{Y|X}$ 是条件累积分布函数。具体而言，训练深度神经网络，使其在 $\alpha_{\text{lo}} = \alpha/2$ 和 $\alpha_{\text{up}} = 1-\alpha/2$处输出下分位数和上分位数 $q^{\alpha_{\text{lo}}}(x)$、$q^{\alpha_{\text{up}}}(x)$ 的估计值。给定一个新的输入 $x^\star$，便可直接构造预测区间 $C_\alpha(x^\star)$，如下所示：

$$C_\alpha(x^\star) = [q_\theta^{\alpha_{\text{lo}}}(x^\star),\ q_\theta^{\alpha_{\text{up}}}(x^\star)]. \tag{3}$$

估计的分位数 $q_\theta^{\alpha_{\text{lo}}}(x^\star)$、$q_\theta^{\alpha_{\text{up}}}(x^\star)$ 可以由单个深度神经网络 f_θ 输出，该网络使用弹珠损失进行训练（Steinwart & Christmann，2011年）。也可以将预测值 $\hat{}y(x^\star)$ 取为 $C_\alpha(x^\star)$ 的中心点。

概率回归 另一种方法是显式地对条件分布 $p(y|x)$ 建模，例如使用高斯模型 $p_\theta(y|x) = \mathcal{N}\left(y;\ \mu_\theta(x)\ \sigma_\theta^2(x)\right)$ (Chua 等人，2018年；Gustafsson 等人，2020a年)。可以训练单个深度神经网络 f_θ，通过最小化相应的负对数似然，同时输出均值 $\mu_\theta(x)$ 和方差 $\sigma_\theta^2(x)$。对于给定的输入 $x^\star$，随后可以构造预测区间，如下所示：

$$C_\alpha(x^\star) = [\mu_\theta(x^\star) - \sigma_\theta(x^\star)\Phi^{-1}(1-\alpha/2),\ \mu_\theta(x^\star) + \sigma_\theta(x^\star)\Phi^{-1}(1-\alpha/2)], \tag{4}$$

其中，Φ 是标准正态分布的累积分布函数。均值 $\mu_\theta(x^\star)$ 也被用作预测值 $\hat{}y$。

认知不确定性 从贝叶斯视角来看，分位数回归和高斯模型可以捕捉偶然不确定性（固有数据噪声），但无法捕捉认知不确定性。认知不确定性反映的是模型参数的不确定性（Gal，2016年；Kendall & Gal，2017年）。通过贝叶斯方法，可以以有原则的方式对其进行估计

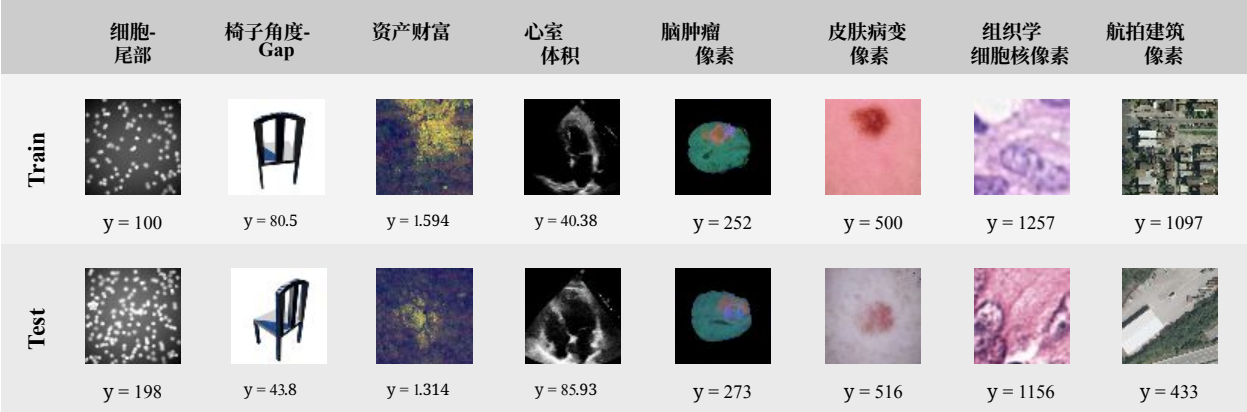


图 1：我们提出了一个由 8 个基于图像的回归数据集构成的基准测试，用于检验回归不确定性估计方法在真实世界分布偏移下的可靠性。此处展示了每个数据集的训练集（上排）和测试输入示例 x，以及相应的真实目标值 y。

推理，并且已有研究探索了各种近似方法（Neal，1995年；Ma 等人，2015年；Blundell 等人，2015年；Gal & Ghahramani，2016年）。实践中，研究表明，很难胜过简单的集成方法（Lakshminarayanan 等人，2017年；Ovadia 等，2019年）；该方法需要训练 M 个模型$\{f_{\theta_i}\}_{i=1}^{M}$，并将它们的预测结合起来。对于高斯模型，可以计算出单个均值 $\hat{}\mu$ 和方差 $\hat{}\sigma^2$，如下所示：

$$\hat{\mu}(x^\star) = \frac{1}{M}\sum_{i=1}^{M}\mu_{\theta_i}(x^\star), \qquad \hat{\sigma}^2(x^\star) = \frac{1}{M}\sum_{i=1}^{M}\left(\left(\hat{\mu}(x^\star) - \mu_{\theta_i}(x^\star)\right)^2 + \sigma^2_{\theta_i}(x^\star)\right), \tag{5}$$

然后代入式 (4)，为给定输入 $x^\star$ 构造预测区间 $C_\alpha(x^\star)$。

2.2 选择性预测

选择性预测框架已应用于分类（Herbei & Wegkamp，2006年；Geifman & El-Yaniv，2017年）和回归问题（Jiang 等，2020年）。其基本思想是让模型可以选择对某些输入不输出预测。实现方法是将预测模型 f_θ 与不确定性函数 κ_f 结合起来：$\mathcal{X} \to \mathbb{R}$。给定输入 $x^\star$，如果不确定性 $\kappa_f(x^\star) \leq \tau$（对于用户指定的某个阈值 τ），则输出预测 $f_\theta(x^\star)$；否则拒绝 $x^\star$，不作预测。预测率 是输出预测的输入所占的比例，

$$\text{Predition Rate} = \frac{1}{N^\star}\sum_{i=1}^{N^\star}\mathbb{I}\{\kappa_f(x_i^\star) \leq \tau\}. \tag{6}$$

原则上，如果高不确定性 $\kappa_f(x^\star)$ 对应较大的预测误差 $|y^\star - \hat{}y(x^\star)|$，反之亦然，那么模型实际输出的所有预测都将具有较小的误差。具体而言，在本研究中，我们将选择性预测与第 2.1 节中的回归方法相结合。因此，当且仅当 (iff) $\kappa_f(x^\star) \leq \tau$ 时，才会输出预测区间$C_\alpha(x^\star)$ 和预测目标 $\hat{}y(x^\star)$。我们的目标是借此改善输出预测区间的校准效果（使区间覆盖率更接近 $1-\alpha$）。

对于不确定性函数 $\kappa_f(x)$，例如可以使用高斯集成 (5) 的方差 $\hat{}\sigma^2(x)$。也可以采用丰富的分布外检测文献中使用的各种不确定性分数（Salehi 等人，2022年）。在分布外检测中，任务是区分分布内输入 x（即与训练集 $\{(x_i,\ y_i)\}_{i=1}^{N}$ 中的输入相似的输入）与分布外输入。一种有理论依据的分布外检测方法，是在训练集上拟合 $p(x)$ 的模型。随后，将满足 $p(x)$ 较小的输入 x 判定为分布外（Serrà 等人，2020年）。在我们考虑的输入 x 为图像的情形下，对 $p(x)$ 建模可能颇具挑战。为缓解这一问题，可以使用特征提取器 $g:\mathcal{X} \to \mathbb{R}^{D_x}$，通过对特征向量 $g(x)$ 拟合一个简单模型，间接对 $p(x)$ 建模。在分类场景中，Mukhoti 等人（2021a）拟合了一个

高斯混合模型（高斯混合模型）到训练集的特征向量 $\{g(x_i)\}_{i=1}^{N}$。给定输入 $x^\star$，如果高斯混合模型密度 GMM$\big(g(x^\star)\big)$ 较小，就将其判定为分布外。Pleiss 等人（2020年）也将这种方法用于回归问题。Sun 等人（2022年；Kuan 和 Mueller，2022年）没有对特征向量拟合高斯混合模型并评估其密度，而是计算 $g(x^\star)$ 与其在训练集 $\{g(x_i)\}_{i=1}^{N}$ 中的 k 个最近邻之间的距离 kNN$\big(g(x^\star)\big)$。如果该 kNN 距离较大，就将输入 $x^\star$ 判定为分布外。

3 拟议基准

我们提出了一个全面的基准测试，用于测试真实世界分布偏移下回归不确定性估计方法的可靠性。该基准测试由 8 个公开可用的基于图像的回归数据集组成，这些数据集将在第3.1节中详细介绍。随后，我们将在第3.2节介绍完整的评估流程，主要根据预测区间覆盖率对不确定性估计方法进行评估。

3.1 数据集

为了建立一个用于分布偏移下基于图像的回归的标准基准测试，我们收集并修改了文献中的 8 个数据集。其中两个包含合成图像，其余六个是真实世界数据集，其中四个来自医疗应用。图 1 展示了这 8 个数据集中的每个数据集的示例。我们为每个合成数据集创建了两个额外的变体，因此总计得到 12 个数据集。它们的规模都相对较大（6 592–20 614 张训练图像），并包含大小为 64×64 的输入图像x 以及 1D 回归目标 y。下面将介绍全部 12 个数据集（附录 A 中还提供了更多细节），首先介绍两个合成数据集及其变体。

细胞给定一张合成荧光显微图像 x，任务是预测图像中的细胞数量 y。我们使用 Ding 等人（2021年；2020年）的 Cell-200 数据集，其中包含 200 000 张尺寸为 64×64 的灰度图像。我们随机抽取 10 000 张训练图像、2 000 张验证图像和 10 000 张测试图像。因此，训练/验证集与测试集之间不存在分布偏移。因此，我们将此数据集用作基线数据集。

细胞-尾部 Cells 的一个变体，训练/验证集与测试集之间存在明显的分布偏移。对于训练/验证集，回归目标 y 被限制在 $]50, 150]$。对于测试集，目标则位于原始范围 $[1, 200]$ 内。

Cells-Gap Cells 的另一个变体，训练/验证集与测试集之间存在明显的分布偏移。对于训练/验证集，回归目标 y 被限制在 $[1, 50[\cup]150, 200]$。对于测试集，目标则位于原始 $[1, 200]$ 内。

椅子角度 给定一张椅子的合成图像 x，任务是预测椅子的偏航角 y。我们使用 RC-49 数据集（Ding 等人，2021年；2020年），其中包含 64×64 张不同椅子模型的图像，这些图像是在偏航角范围为 $0.1°$ 至 $89.9°$的条件下渲染的。我们随机划分其训练集，得到 17 640 张训练图像和 4 410 张验证图像。通过对其测试集进行子采样，我们还获得了 11 225 张测试图像。训练/验证集与测试集之间没有明显的分布偏移，因此我们将其用作第二个基线数据集。

椅子角度-尾部 这是椅子角度的一个变体，训练/验证集与测试集之间存在明显的分布偏移。对于训练/验证集，我们将回归目标 y 限制为 $]15, 75[$。对于测试集，目标则位于原始 $]0, 90[$范围内。

椅子角度-间隙 这是椅子角度的另一个变体，存在明显的分布偏移。对于训练/验证集，回归目标 y 被限制为 $]0$、$30[\cup]60$、$90[$。对于测试集，目标则位于原始 $]0, 90[$范围内。

资产财富给定一张卫星图像 x，任务是预测该区域的资产财富指数 y。我们使用来自 Koh 等人（2021）的 PovertyMap-Wilds 数据集，该数据集是 Yeh 等人（2020年）收集的数据集的一个变体。我们使用数据中的训练集、ID 验证集和 OOD 测试集子集，分别包含 9 797 张训练图像、1 000 张验证图像和 3 963 张测试图像。我们将图像尺寸从 224×224 调整为 64×64。训练/验证集和测试集包含来自不同非洲国家的卫星图像，因此产生了分布偏移。

心室容积 给定一张人类心脏的超声心动图图像 x，任务是预测左心室的容积y。我们使用 EchoNet-Dynamic 数据集（Ouyang 等人，2020年），其中包含 10 030 段超声心动图视频。每段视频记录一个完整的心动周期，并标注两个不同时间点的左心室容积，分别对应收缩末期（收缩结束时）和舒张末期（收缩开始前）。对于每段视频，我们仅提取其中一个容积测量值及其对应的视频帧。为了在训练/验证集与测试集之间形成明显的分布偏移，我们选择收缩末期的

容积（较小的容积）作为训练集和验证集的数据，而选择舒张末期容积（较大的容积）作为测试集的数据。我们使用提供的数据集划分，得到 7 460 张训练图像、1 288 张验证图像和 1 276 张测试图像。

BrainTumourPixels 给定一张脑部 MRI 扫描的图像切片 x，任务是预测图像中被标记为脑肿瘤的像素数量 y。我们采用医学分割十项全能挑战赛的脑肿瘤数据集（Simpson 等人，2019年；Antonelli 等人，2022年），该数据集是2016年和2017年 BraTS 挑战赛所使用数据的一部分（Bakas 等人，2018年；2017年；Menze 等人，2014年）。该数据集包含484个脑部 MRI 扫描及其对应的肿瘤分割掩膜。我们按照80%/20%/20%的比例，将这些扫描划分为训练集、验证集和测试集。这些扫描是大小为 $240 \times 240 \times 155$ 的三维体积。我们将每次扫描转换为155张大小为 240×240 的图像切片，并通过统计每张图像中标记为脑肿瘤的像素数量，为每张图像创建一个回归目标。由此得到20 614张训练图像、6 116张验证图像和6 252张测试图像。

皮肤病变像素 给定一张色素性皮肤病变的皮肤镜图像 x，任务是预测图像中被标记为病变的像素数量 y。我们使用 Tschandl 等人（2018年）的 HAM10000 数据集，其中包含 10 015 张皮肤镜图像及其对应的皮肤病变分割掩膜。HAM10000 由四个不同的子数据集组成，其中三个采集于奥地利，第四个采集于澳大利亚。为了在训练/验证集与测试集之间形成明显的分布偏移，我们将澳大利亚子数据集用作测试集。将剩余图像随机按 85%/15% 划分为训练集和验证集后，我们得到 6 592 张训练图像、1 164 张验证图像和 2 259 张测试图像。

HistologyNucleiPixels 给定一张苏木精-伊红染色组织学图像 x，任务是预测图像中被标记为细胞核的像素数量 y。我们使用 Graham 等人（2019年）的 CoNSeP 数据集，以及他们提供的 Kumar 数据集（Kumar 等人，2017年）和 TNBC 数据集（Naylor 等人，2018年）的预处理版本。这些数据集包含大型苏木精-伊红染色图像块及其对应的细胞核分割掩膜。这三个数据集采集自不同的医院/机构，标本保存和染色流程也各不相同。因此，使用 CoNSeP 和 Kumar 数据集作为训练/验证集，并使用 TNBC 作为测试集，可以形成明显的分布偏移。我们通过规则网格划分，从大型图像块中提取 64×64 图像切片，并通过统计每个图像切片中被标记的细胞核像素数量，为其创建回归目标。最终，我们得到 10 808 张训练图像、2 702 张验证图像和 2 267 张测试图像。

AerialBuildingPixels 给定一张航拍图像 x，任务是预测图像中被标注为建筑物的像素数量 y。我们使用 Inria 航拍图像标注数据集（ Maggiori 等人，2017 年），该数据集包含 180 张大型航拍图像及其对应的建筑物分割掩码。这些图像采集自五个不同的地理位置。我们使用来自两座人口稠密的美国城市的图像作为训练/验证集，并使用来自欧洲较乡村地区的图像作为测试集，从而形成明显的分布偏移。经过预处理后，我们得到 11 184 张训练图像、2 797 张验证图像和 3 890 张测试图像。

构建这些定制数据集是我们的主要贡献之一。这使我们能够针对基于图像的回归，提出一个涵盖不同类型挑战性分布偏移的大规模且易于使用的数据集基准测试（这些数据集均公开可用）。

3.2 评估

我们提出，主要根据预测区间覆盖率 (1) 来评估回归不确定性估计方法：对于每个输入 x，如果一种方法输出预测值 $\hat{y}(x)$ 和一个 90% 的预测区间 $C_{0.1}(x)$，那么该方法在测试集 上是否确实达到了 90% 的覆盖率？也就是说，预测区间是否经过校准?

动机回归不确定性估计方法可以通过多种方式进行评估。一种替代方法是稀疏化（Ilg 等人，2018），用于衡量不确定性在将预测从最差到最好排序时的效果。然而，即使不确定性的绝对尺度始终被低估，也仍然可以实现完美的稀疏化。因此，许多先前的工作转而从校准角度评估方法（Gneiting等，2007；Guo 等人，2017）。对于回归，常见的一种校准形式基于分位数（Kuleshov 等人，2018；Cui 等人，2020；Chung 等人，2021）。本质上，如果区间覆盖率 (1) 对所有失覆盖率 $\alpha \in]0,1[$都等于 $1-\alpha$，就称模型经过了校准。该指标通过期望校准误差 ECE $= \frac{1}{m}\sum_{j=1}^{m} \left|\text{Coverage}(C_{\alpha_j}) - (1-\alpha_j)\right|$, $\alpha_j \sim \text{U}(0, 1)$进行衡量。因此，我们提出的评估指标是这一方法的一个特例，只考虑一种特定的失覆盖率 $\alpha = 0.1$。我们认为，这样得到的指标更简单、也更易于解释，同时也符合预测区间在现实应用中的实际使用方式

：在实际应用中，人们会选定 α 的一个特定值（通常选择 $\alpha = 0.1$ ），然后期望相应的预测区间 C_α 在未见过的测试数据上达到 $1-\alpha$ 的覆盖率。近期提出的其他校准指标则直接衡量不确定性与预测误差的吻合程度（Levi等，2022；Pickering与Sapsis，2022）。这些指标虽然能够对不同方法进行相对比较，但难以从绝对性能的角度加以解释。

实现细节 对于每个数据集和方法，我们首先在训练集$\mathcal{D}_{\text{train}}$上训练一个深度神经网络。然后，在验证集 $_{\text{val}}$上运行该方法，为每个输入 x ($\alpha = 0.1$)生成一个预测区间 $C_\alpha(x) = [L_\alpha(x) \;\; U_\alpha(x)]$ 。重要的是，接着我们按照˜ (Romano等，2019)中的流程，在验证集 上对这些预测区间进行校准。由此得到校准的预测区间 $C_\alpha(x)$，其在验证集˜上的区间覆盖率恰好等于 $1-\alpha$。具体而言， $C_\alpha(x)$由原始区间 $C_\alpha(x) = [L_\alpha(x) \;\; U_\alpha(x)]$,构造而成。

$$\tilde{C}_\alpha(x) = [L_\alpha(x) - Q_{1-\alpha}(E, \mathcal{D}_{\text{val}}),\; U_\alpha(x) + Q_{1-\alpha}(E, \mathcal{D}_{\text{val}})], \tag{7}$$

其中， $E = \{\max\big(L_\alpha(x_i) - y_i,\, y_i - U_\alpha(x_i)\big) : i \in \mathrm{D}_{\text{val}}\}$ 是在 D_{val} 上计算的一致性分数， $Q_{1-\alpha}(E, \mathrm{D}_{\text{val}})$ 是这些分数的第 $(1-\alpha)$ 个分位数。最后，我们在测试集 D_{test} 上运行该方法，为每个输入 $x^\star$输出一个校准的预测区间 $C_\alpha(x^\star)$。理想情况下， $C_\alpha(x^\star)$ 的区间覆盖率从验证集到测试集不会˜ ˜改变，即在测试集上 Coverage(C_α) $= 1-\alpha$ 也应成立。如果 Coverage(C_α) $\neq 1-\alpha$，则与过度自信方法 ($< 1-\alpha$) 相比，更倾向于采用保守方法 ($> 1-\alpha$)。对于˜基于选择性预测的方法（第 2.2 节），唯一的区别在于，预测区间 $C_\alpha(x^\star)$ 仅针对部分测试输入 $x^\star$ 输出（当且仅当 $\kappa_f(x^\star) \leq \tau$）。因此，区间覆盖率仅在这部分测试集上计算。这与 Fisch等（2022）在分类场景中讨论的“选择性校准”概念类似。我们设定 $\alpha = 0.1$ ，因为这是实践中常用的失覆盖率。

次级指标 我们还根据验证集上的平均绝对误差（MAE）和平均区间长度来评估方法。这两项指标分别衡量预测 $\hat{}y(x)$ 和预测区间 $C_\alpha(x)$ 的质量（Gneiting等，2007）。平均区间长度是一项自然的次级指标，因为如果某种方法虽然实现了接近 $1-\alpha$ 的覆盖率，却对所有输入 x 都输出极大的区间，那么它在实践中并没有多大用处。此外，如果两种不同的方法都经过完美校准，即 Coverage(C_α) $= 1-\alpha$，则应优先选择生成较小预测区间的方法。对于基于选择性预测的方法（仅针对某些输入 x 输出预测），实际输出预测的输入所占比例也是一项自然的次级指标。因此，我们还通过测试集上的预测率（6）来评估这些方法。如果覆盖率接近 $1-\alpha$，但预测率很低，那么这种方法在某些应用中仍可能不实用。对于两种经过完美校准的方法，应优先选择预测率较高的方法。

4 评估方法

我们评估了第 2.1 节中的五种常见回归不确定性估计方法，这些方法都会为每个输入 x 输出一个预测区间 $C_\alpha(x)$ 和一个预测目标 $\hat{}y(x) \in C_\alpha(x)$。其中两种方法还与选择性预测（第 2.2 节）结合，并使用四种不同的不确定性函数 $\kappa_f(x)$。总共评估了 10 种不同的方法。对于所有方法，我们都基于 ResNet34（He 等人，2016年）骨干深度神经网络训练模型。选择这种架构是因为它简单且应用广泛。ResNet 以图像 x 作为输入，并输出一个特征向量 $g(x) \in \mathbb{R}^{512}$。下文将说明这 10 种评估方法，并提供各自的实现细节；更一般的介绍请参见第 2 节。

保形预测 我们将 ResNet 特征向量$g(x)$ 输入由两个全连接层组成的网络头，从而创建一个标准的直接回归模型，并输出一个标量预测值 $f_\theta(x)$。该模型使用 L2 损失进行训练。随后，我们根据公式（2）使用保形预测来构造预测区间 $C_\alpha(x)$。我们没有将训练图像划分为 $\mathcal{I}_1$ 和 $\mathcal{I}_2$, ，而是将验证图像用作校准集 $\mathcal{I}_2$。

集成 我们训练由 $M = 5$ 个直接回归模型组成的集成$\{f_{\theta_1}, \ldots, f_{\theta_M}\}$，并计算集成均值 $\hat{}\mu(x) = \frac{1}{M}\sum_{i=1}^{M} f_{\theta_i}(x)$ 和集成方差 $\hat{}\sigma^2(x) = \frac{1}{M}\sum_{i=1}^{M}\big(\hat{}\mu(x) - f_{\theta_i}(x)\big)^2$。将这些值代入公式 4，即可构造长度依赖于输入的预测区间 $C_\alpha(x)$。

高斯 我们通过将 ResNet 特征向量$g(x)$ 输入两个独立的网络头（每个网络头均由两个全连接层组成），来创建高斯模型 $p_\theta(y|x) = \mathcal{N}\left(y;\ \mu_\theta(x)\ \ \sigma_\theta^2(x)\right)$。这两个网络头分别输出均值 $\mu_\theta(x)$ 和方差$\sigma_\theta^2(x)$。随后根据 (4) 构造预测区间 $C_\alpha(x)$。

高斯集成 我们训练一个由 $M = 5$ 个高斯模型组成的集成，根据（5）计算单个均值 $\hat{}\mu(x)$ 和方差 $\hat{}\sigma^2(x)$，并将它们代入（4）以构造预测区间 $C_\alpha(x)$。

分位数回归 我们将 ResNet特征向量 $g(x)$ 输入两个独立的网络头，从而创建一个分位数回归模型。这两个网络头输出分位数 $q_\theta^{\alpha_{\text{lo}}}(x)$ $q_\theta^{\alpha_{\text{up}}}(x)$，并根据（3）直接构成预测区间 $C_\alpha(x)$。模型通过最小化 $q_\theta^{\alpha_{\text{lo}}}(x)$ 和 $q_\theta^{\alpha_{\text{up}}}(x)$ 的平均弹珠损失进行训练。

高斯 + 选择性高斯混合模型 我们将高斯方法与选择性预测机制相结合。训练好高斯模型后，我们将其应用于训练集中的每张图像，以提取 ResNet特征向量 $\{g(x_i)\}_{i=1}^N$。随后，我们利用 Scikit-learn（Pedregosa 等人，2011年）对这些训练特征向量拟合一个高斯混合模型（4 个分量，全协方差）。对于给定的输入 x，为了计算不确定性分数 $\kappa_f(x)$，我们提取 $g(x)$，并根据拟合后的高斯混合模型评估其似然度，$\kappa_f(x) = -\text{高斯混合模型}\big(g(x)\big)$。仅当 $\kappa_f(x) \leq \tau$ 时，才输出高斯模型的预测 $\mu_\theta(x)$ 及相应的预测区间 $C_\alpha(x)$。为设定用户指定的阈值 τ，我们计算验证集所有图像上的 $\kappa_f(x)$，并选取第 95 个分位数。选择 τ 是出于常用的 FPR95 分布外检测指标的考虑，但 τ 也可以通过其他方法设定。

高斯 + 选择性 k 近邻 与高斯 + 选择性高斯混合模型相同，但采用 $\kappa_f(x) = \text{kNN}\big(g(x)\big)$。具体而言，提取 $g(x)$ 后，计算其与训练特征向量中 $k = 10$ 个最近邻的平均距离，以此得到不确定性分数 $\kappa_f(x)$。$\{g(x_i)\}_{i=1}^N$。遵循 Kuan 和 Mueller（2022年）的做法，我们使用 Annoy[1] 近似近邻库，并将余弦相似度作为距离度量。

高斯 +选择性方差 与高斯 +选择性高斯混合模型相同，但采用 $\kappa_f(x) = \sigma_\theta^2(x)$（高斯模型的方差）。这被用作前两种方法的简单基线。

高斯集成 + 选择性高斯混合模型 我们将高斯集成与选择性预测机制相结合。训练由 $M = 5$ 个高斯模型组成的集成后，我们让每个模型处理训练集中的每张图像，以提取 M 组 ResNet特征向量。随后，我们为每个模型的训练特征向量集合拟合一个高斯混合模型。也就是说，我们拟合 M 个不同的高斯混合模型。要计算给定输入 x的不确定性分数 $\kappa_f(x)$，我们提取一个特征向量，并针对 M 个模型中的每个模型，根据相应的高斯混合模型评估其似然度。最后，我们计算高斯混合模型似然度的均值，$\kappa_f(x) = \frac{1}{M}\sum_{i=1}^M -\text{GMM}_i\big(g_i(x)\big)$。

高斯集成 +选择性集成方差 与高斯集成 +选择性高斯混合模型相同，但 $\kappa_f(x) = \frac{1}{M}\sum_{i=1}^M \left(\hat{}\mu(x) - \mu_{\theta_i}(x)\right)^2$，其中 $\hat{}\mu(x) = \frac{1}{M}\sum_{i=1}^M \mu_{\theta_i}(x)$。因此，集成均值的方差被用作不确定性分数。这构成了前一种方法的简单基线。

所有模型均使用 ADAM优化器（Kingma与Ba，2014年）训练 75 个训练轮次。所有数据集使用相同的超参数，并未针对任何特定数据集对训练流程或模型进行专门调优。所有实验均使用 PyTorch（Paszke等人，2019年）实现，且我们已公开完整的实现代码。所有模型均在单块 NVIDIA TITAN Xp GPU 上训练。在一块此类 GPU 上，针对一个数据集训练 20 个模型约需 24 小时。我们选择基于 ResNet34骨干网络训练所有模型，因为 ResNet34广泛应用于各种场景，同时易于实现且计算成本较低。然而，拟议基准测试和评估的不确定性估计方法完全不依赖于这一特定的深度神经网络骨干网络架构。探索使用其他更强大的模型，并评估其对不确定性估计方法可靠性的影响，是一个有趣的研究方向，我们将留待未来工作。

5 相关工作

深度神经网络的分布外鲁棒性是一个活跃的研究领域（Hendrycks & Dietterich, 2019年；Hendrycks 等人, 2021年b;a；Izmailov等人, 2021；Yao等人, 2022年；Schwinn等人, 2022年；Wiles等人, 2022年；Zhang等人, 2022年）。然而，所有这些先前的研究都只关注分类任务。此外，它们并未考虑任何不确定性度量，而只根据准确率进行评估。虽然（Zaidi等人, 2021；Hendrycks

[1] https://github.com/spotify/annoy

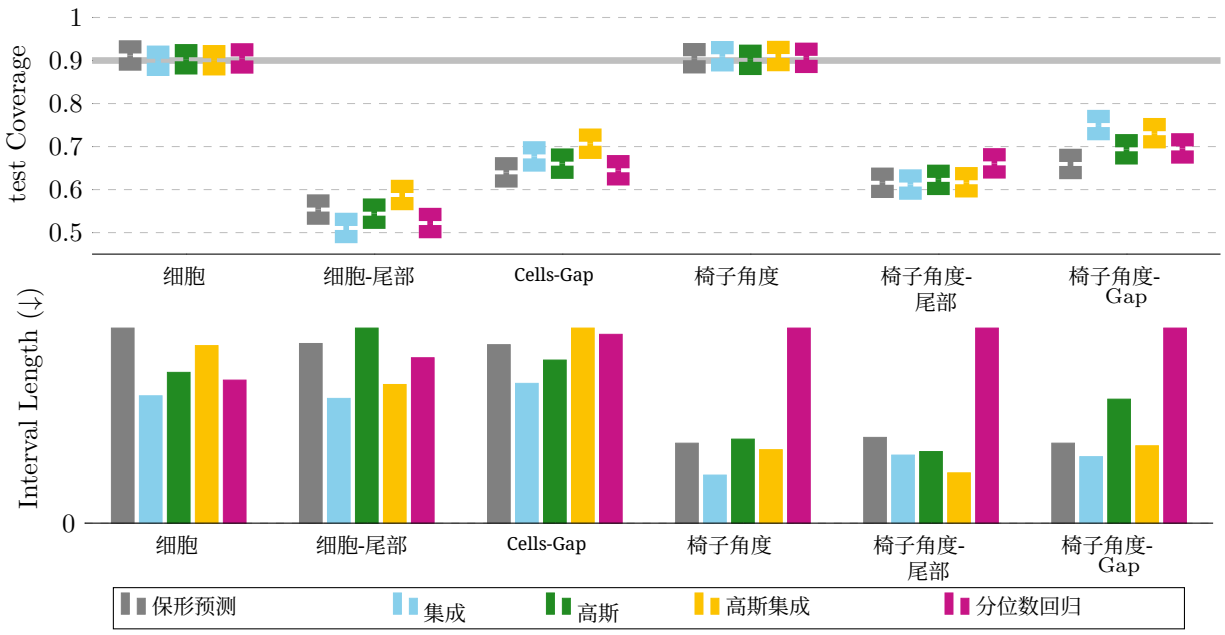


图2：五种常见回归不确定性估计方法（这些方法会针对所有输入输出预测及对应的 90% 预测区间）在六个合成数据集上的结果。**上图：**以我们的主要指标——测试集覆盖率衡量的结果。完美校准的方法应达到恰好 90% 的测试集覆盖率，如实线所示。**下图：**以平均验证区间长度衡量的结果。

等人, 2022）评估了不确定性校准，但也只考虑分类场景。相比之下，评估不确定性估计方法是我们的主要关注点，而我们具体针对回归开展这项工作。

我们的工作主要受到 (Koh 等，2021) 和 (Ovadia 等，2019) 的启发。Koh 等人（2021）提出了一个涵盖各种真实世界分布偏移的大型基准测试，但其中只包含一个回归数据集。此外，评估仅依据预测性能进行。Ovadia 等人（2019）对分布偏移下的不确定性估计方法进行了全面评估，但只考虑了分类任务。受此启发，我们提出了这一基准测试，用于评估不确定性估计方法在真实世界分布偏移下的回归可靠性。与我们最相似的是 Malinin 等人（2021）的工作。然而，他们的基准测试仅包含两个回归数据集（表格天气预测和一项复杂的车辆运动预测任务），只评估基于集成的不确定性估计方法，而且没有依据校准情况评估这些方法。

6 结果

我们按照第3.2节所述的评估流程，对第3.1节中的全部12个数据集评估第4节指定的10种方法。对于每种方法，我们训练20个模型，随机选取其中5个进行评估，并报告平均指标。对于集成方法，我们从20个训练好的模型中随机选取 $M = 5$ 个来构建集成，对该集成进行评估，并总共重复5次。为确保结果不依赖于我们对 $\alpha = 0.1$的具体选择，我们还使用另外两种失覆盖率评估这些方法。正文仅包含 $\alpha = 0.1$的结果；我们还在附录B中对 $\alpha = 0.2$ 和 $\alpha = 0.05$ 进行了大部分评估，整体观察到的趋势非常相似。

6.1 常见不确定性估计方法

我们首先评估第 4 节中的前五种方法，这些方法会针对所有输入输出预测及相应的 90% 预测区间。合成数据集的主要指标——测试集覆盖率——结果见图2上半部分；六个真实世界数据集的结果见图3。在

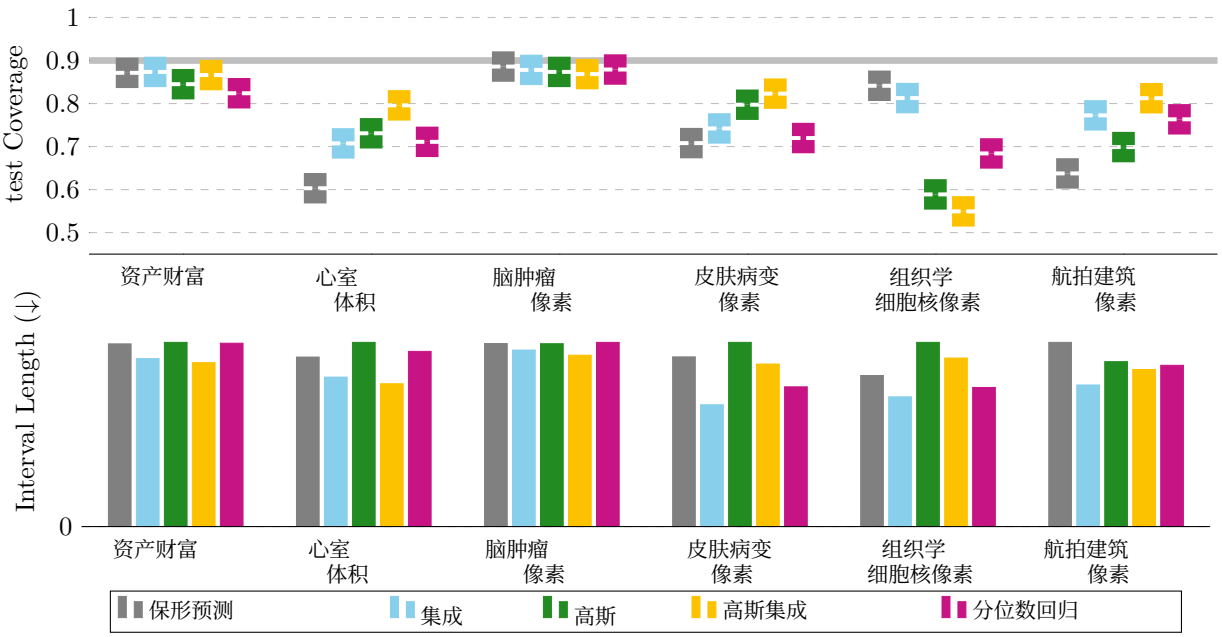


图3：五种常见回归不确定性估计方法（针对所有输入输出预测及相应的90%预测区间）在六个真实世界数据集上的结果。**上图：**以主要指标测试集覆盖率衡量的结果。**下图：**以平均验证区间长度衡量的结果。

图2和图3的下半部分展示了平均验证区间长度方面的结果。完整结果（包括另一项次要指标——验证集平均绝对误差）见附录中的表A1至表A12。请注意，由于我们使用的是一个由自定义数据集组成的新基准测试，因此无法将模型的平均绝对误差与文献中先前研究的结果直接比较。

图2左上方展示了第一个合成数据集细胞上的测试集覆盖率结果。由于该数据集的训练/验证集与测试集之间不存在分布偏移，因此我们将其用作基线。我们观察到，五种方法的预测区间都几乎经过了完美校准，即它们获得的测试集覆盖率都非常接近90%。这正是我们期望的表现。然而，在引入明显分布偏移的Cells-Tails数据集上，我们从图2中观察到，五种方法的测试集覆盖率都大幅偏离了期望的90%。即使是最先进的不确定性估计方法高斯集成也变得过度自信，其测试集覆盖率降至 $\approx 59\%$。在Cells-Gap数据集上，测试集覆盖率稍微接近90%，但五种方法仍然都过度自信。在另一个合成数据集ChairAngle上，我们从图2中观察到，五种方法的预测区间都几乎经过了完美校准。然而，在ChairAngle-Tails和ChairAngle-Gap上引入明显的分布偏移后，我们可以看到，所有方法的测试集覆盖率再次大幅下降。

六个真实世界数据集上的结果见图3。上半部分显示，尽管五种方法的表现始终略显过度自信（测试集覆盖率为82%–89%），但它们在资产财富和 BrainTumourPixels 上的预测区间校准得相当好。在其余四个数据集上，这些方法通常过度自信的程度更为明显。在心室容积数据集上，所有方法的测试集覆盖率为60%–80%；在皮肤病变像素数据集上，最高覆盖率为 $\approx 82\%$。在 HistologyNucleiPixels 上，大多数方法的测试集覆盖率仅为55%–70%；在 AerialBuildingPixels 上，最高覆盖率为 $\approx 81\%$。事实上，在这些真实世界数据集上，没有一种方法真正达到预期的90%测试集覆盖率。

从平均验证区间长度来看，我们在图2和图3下半部分观察到，集成方法生成的预测区间始终比保形预测更短。此外，高斯集成的区间通常比高斯方法的更短。比较分位数回归和高斯方法的区间长度时，我们没有发现适用于所有数据集的一致趋势。

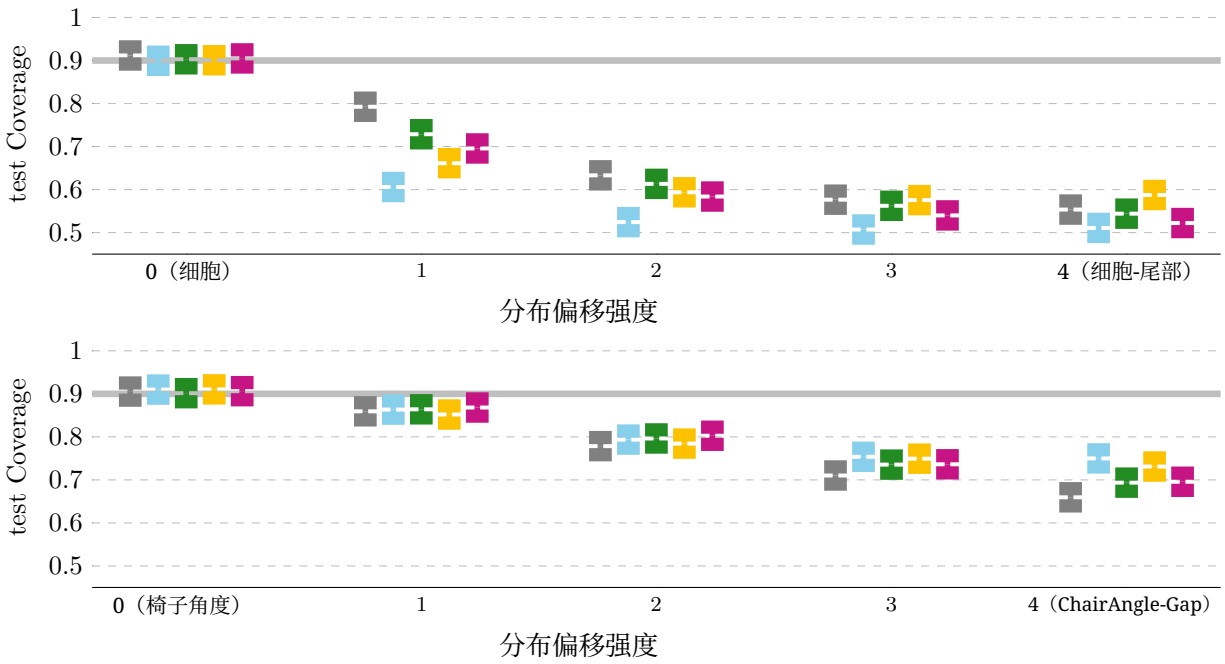


图 4：五种常见回归不确定性估计方法在分布偏移程度逐渐增加的合成数据集上的测试覆盖率结果。分布偏移程度逐渐增加。**上图：**从 Cells（无分布偏移）到 Cells- Tails（最大分布偏移）。**下图：**从 ChairAngle 到 ChairAngle-Gap。

由于不同数据集之间的平均区间长度差异很大，图2和图3仅展示各方法的相对比较。绝对数值尺度请参见表A1至表A12。

为了进一步研究分布偏移如何影响测试覆盖率性能，我们还将这五种方法应用于 Cells 数据集的另外三种变体。Cells 的训练/验证集与测试集的回归目标范围没有差异，而 Cells-Tails 的训练/验证集目标范围为 $]50, 150]$ ，测试集目标范围为$[1, 200]$ 。通过创建三个目标范围介于两者之间的变体，我们由此得到一组包含五个数据集的序列，其分布偏移程度逐渐增加，从 Cells（无分布偏移）开始，到 Cells-Tails（最大分布偏移）结束。图 4 上半部分展示了这组数据集上的测试覆盖率结果。我们观察到，随着分布偏移程度逐步增加，测试覆盖率也随之下降。图 4 下半部分展示了类似实验的结果：我们构建了一组包含五个数据集的序列，从 ChairAngle（无分布偏移）开始，到 ChairAngle-Gap（最大分布偏移）结束。在这种情况下，我们同样观察到，随着分布偏移程度逐步增加，测试覆盖率也逐步下降。

附录中的图 A1 至图A3最终展示了在三个指标（测试集覆盖率、平均验证区间长度、验证集平均绝对误差）方面，五种方法在真实世界数据集上的相对性能研究。可以清楚地观察到，集成和高斯集成取得了最佳性能，这表明集成多个模型确实有助于提升性能。

6.2 选择性预测方法

接下来，我们评估加入选择性预测机制后的各方法。首先考察三种基于高斯的方法。合成数据集上的测试集覆盖率和测试集预测率结果见图 5，六个真实世界数据集上的结果见图 6。虽然完整评估这些方法也应包括平均验证区间长度，但我们注意到，选择性预测机制不会改变底层高斯方法的区间（该方法的区间长度已在第6.1节中评估）。因此，我们在此重点关注测试集覆盖率和测试集预测率。完整的数值结果见附录中的表A1至表A12。

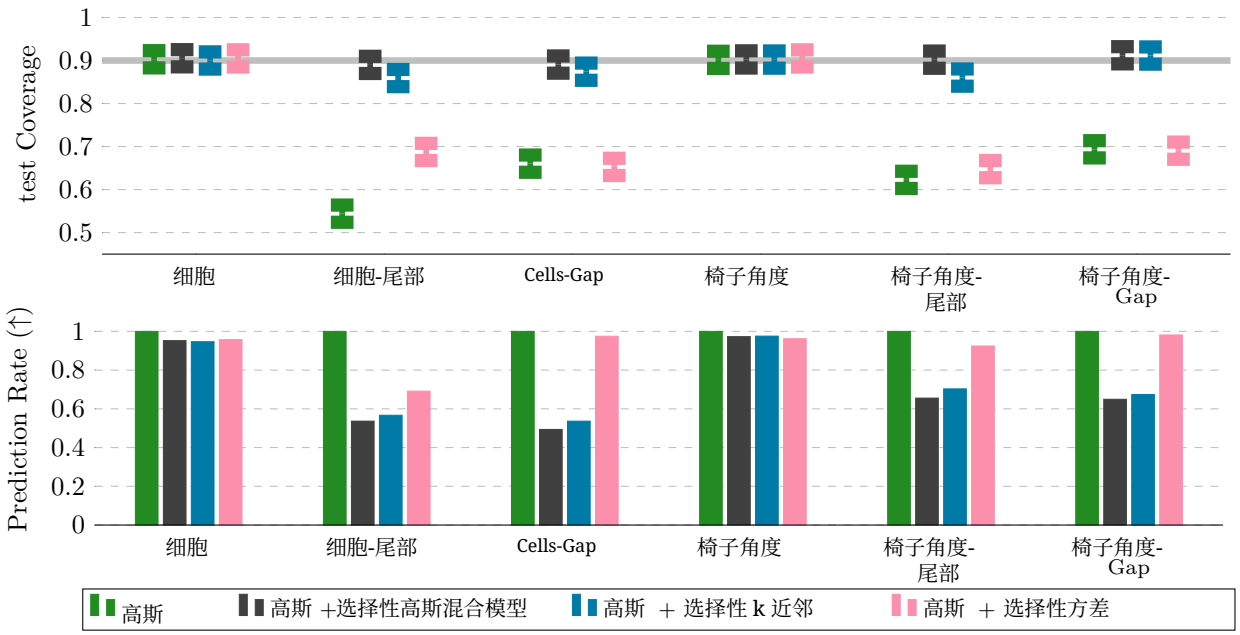


图 5：基于高斯的三种选择性预测方法在六个合成数据集上的结果。**上：** 以主要指标测试集覆盖率衡量的结果。**下：** 以测试集预测率（实际输出预测的测试输入所占比例）衡量的结果。

从图 5 上半部分可以看出，基于特征空间密度的选择性预测显著提高了高斯方法在合成数据集上的测试集覆盖率。高斯方法仅在细胞和椅子角度这两个没有任何分布偏移的基线数据集上具有校准良好的预测区间，而高斯 + 选择性高斯混合模型在全部六个数据集上几乎都实现了完美校准。例如，在细胞-尾部数据集上，其测试集覆盖率从 $\approx 54\%$ 提高到 $\approx 89\%$。高斯 + 选择性 k 近邻也显著提高了测试集覆盖率，但提升幅度略逊一筹。从图 5 下半部分可以看出，当高斯 + 选择性高斯混合模型显著提高测试集覆盖率时，其测试集预测率也明显下降。例如，在细胞-尾部数据集上，预测率从细胞数据集上的 ≈ 0.95 降至 ≈ 0.54 。在这种情况下，通过将近 50% 的输入判定为分布外并予以拒绝，高斯 + 选择性高斯混合模型因此能够在其实际输出预测的测试集子集上保持良好的校准。在图 5 中还可以看到，高斯 + 选择性方差仅略微提高了测试集覆盖率。

虽然高斯 + 选择性高斯混合模型显著提高了高斯方法的测试集覆盖率，并且在合成数据集上具有校准良好的预测区间，但我们从图 6 中观察到，这一结论并不适用于六个真实世界数据集。高斯 + 选择性高斯混合模型确实始终能提高测试集覆盖率，但提升幅度很小，而且在许多情况下仍存在显著的过度自信。例如，在心室容积数据集上，高斯 + 选择性高斯混合模型的测试集预测率低至 ≈ 0.71，但与高斯方法相比，测试集覆盖率仅从 $\approx 73\%$ 提高到 $\approx 75\%$。

基于高斯集成的两种方法的结果见附录中的图 A4 和图 A5。总体而言，我们观察到非常相似的趋势。高斯集成 + 选择性高斯混合模型显著提高了高斯集成的测试集覆盖率，并且在合成数据集上几乎完全校准。然而，在真实世界数据集上，它往往仍存在显著的过度自信。

最后，图 A6 对比了 5 种选择性预测方法在全部 12 个数据集上的相对表现，比较指标为平均测试覆盖误差（经验区间覆盖率与预期区间覆盖率之间的绝对差值）和平均测试预测率。我们观察到，高斯集成 + 选择性高斯混合模型的测试覆盖误差最低，但测试预测率也最低。事实上，对于这 5 种方法，测试覆盖误差每有改善，测试预测率也会随之下降，这意味着这两个指标之间似乎存在固有的权衡。

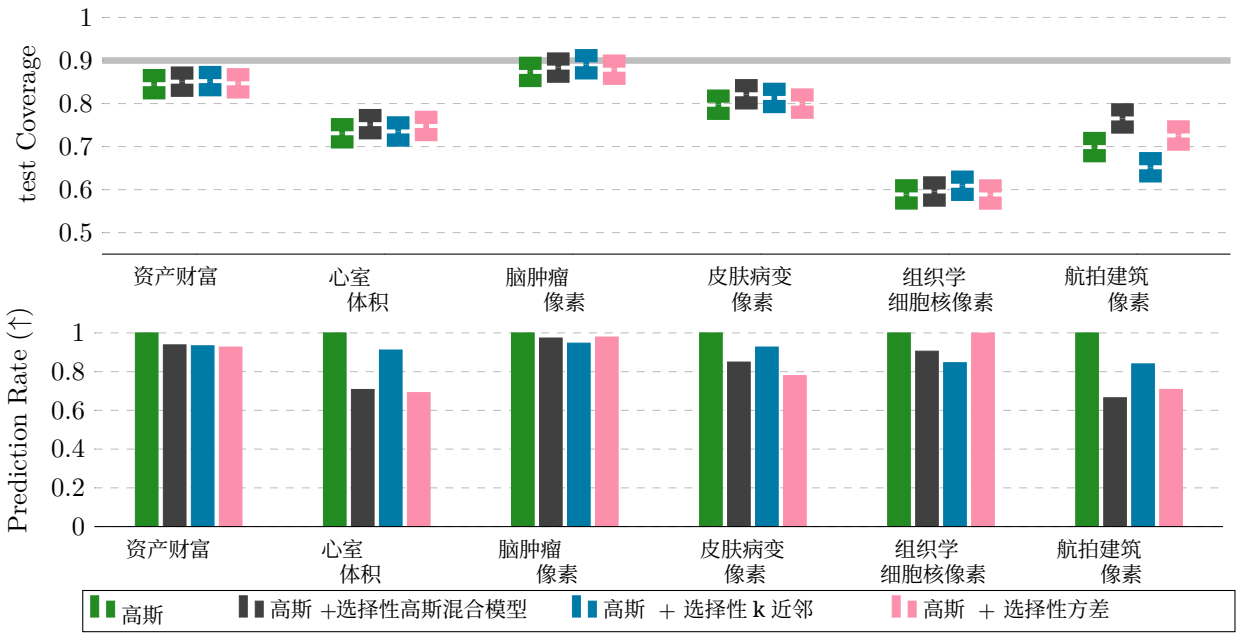


图 6：基于高斯的三种选择性预测方法在六个真实世界 数据集上的结果。**上：** 测试集覆盖率方面的结果。**下：** 测试集预测率方面的结果。

7 讨论

现在，让我们更详细地分析第 6 节的结果，并讨论我们认为最重要的发现和见解。首先可以看到，在考察的 10 种方法中，没有一种能在全部 12 个数据集上生成接近完美校准的预测区间。因此，我们认为拟议基准确实具有挑战性，也很有意义。此外，图2和图3中的结果表明，在不存在分布偏移时（细胞和椅子角度），常见的不确定性估计方法能够得到良好校准；但在许多现实场景中，这些方法都会失效，变得过度自信。这凸显了在评估不确定性估计方法时，采用足够贴近现实且具有挑战性的基准测试的重要性。否则，我们可能会误以为这些方法在实际部署中的可靠性高于其真实水平。

覆盖率保证可能会带来虚假的安全感 我们还想强调，保形预测和分位数回归[2] 具有理论覆盖率保证，但从图2和图3中可以观察到，它们在许多数据集上仍会变得过度自信。由于这些保证依赖于所有数据点均可交换这一假设（例如，独立同分布数据满足该假设），而分布偏移下通常并不满足这一条件，因此这些结果实际上并不令人意外。不过，这些结果很好地提醒我们，始终需要留意方法所依赖的假设，以及这些假设在常见实际应用中是否可能成立。否则，这些理论保证可能只会带来虚假的安全感，让我们对方法的信任超过实际应有的程度。

合成数据集与真实世界数据集之间明显的性能差异 我们发现一个有趣的现象：基于特征空间密度的选择性预测，尤其是高斯 + 选择性高斯混合模型，在合成数据集上几乎能够完美实现测试集覆盖率（图 5），但在真实世界数据集上却未能带来显著改进（图 6）。心室容积上的结果尤其值得关注，因为预测率大幅下降，测试集覆盖率却没有显著提升。这意味着，尽管相当大比例的测试输入被判定为分布外，因而被该方法拒绝，测试集覆盖率却几乎没有改善。在合成数据集上，每当预测率显著下降时，测试集覆盖率也会相应提高（图 5）。目前尚不清楚

[2]由于所有预测区间都是在验证集上校准的，因此我们采用保形化分位数回归 （Romano等，2019）。

为何合成数据集与真实世界数据集之间会出现如此明显的性能差异。一种可能的解释是，真实世界数据需要更好的模型来处理 $p(x)$，也就是说，基于特征空间密度的相对简单方法并不足够。恰当地解释这种性能差异是一个重要问题，但我们在此将其留作未来研究的一个有趣方向。

分布外不确定性分数表现良好，但仍不够理想比较各种选择性预测方法后，我们发现，高斯 + 选择性高斯混合模型始终优于高斯 + 选择性方差（图 5 和图 6）；在大多数情况下，高斯集成 + 选择性高斯混合模型也优于高斯集成 + 选择性集成方差（图 A4 和图 A5）。相较于常见的不确定性估计基线，基于特征空间密度的方法表现非常出色。这与近期展示出的最先进分布外检测性能一致。不过，从我们的结果中可以清楚地看到，尽管基于特征空间的方法相对常见基线表现良好，其产生的选择性预测方法在许多情况下仍然过度自信——其绝对性能仍远非完美。因此，借助我们的基准测试，我们不仅能够比较不同分布外不确定性分数的相对性能，还能从绝对角度评估其性能。

真实世界数据集之间的性能差异大多合乎逻辑 比较图3中不同真实世界数据集上的性能时，我们发现所有方法在BrainTumourPixels和AssetWealth上的校准效果都相对较好。对于BrainTumourPixels，训练集、验证集和测试集是通过随机划分原始磁共振成像扫描集合创建的。因此，训练/验证集与测试集之间的分布偏移也相当有限。对于AssetWealth（来自不同非洲国家的卫星图像），与AerialBuildingPixels（来自两个不同大洲的卫星图像）相比，其分布偏移可能相当有限。最后，HistologyNucleiPixels的结果很有意思，因为这是唯一一个保形预测明显取得最佳测试集覆盖率的数据集。目前尚不清楚为何输出长度取决于输入的预测区间的方法在这个特定数据集上表现不佳。

最后，我们应当强调，尽管测试集覆盖率是我们的主要指标，但仅凭这一指标并不足以说明某种方法总体上“表现良好”。例如，测试集预测率较低的完美校准方法，在实践中可能并无特别大的用处。虽然在许多医疗应用和其他安全关键领域，即使测试集预测率非常低也很可能可以接受（只要方法始终保持完美校准），但可以想象，在风险较低的场景中，校准与预测率之间的这种权衡就没有那么明确。由于所评估的方法中，没有一种能在全部12个数据集上接近完美校准，因此本文主要着重分析测试集覆盖率。如果有多种方法在测试集覆盖率方面表现良好，我们就有必要更详细地分析和讨论次级指标的表现。

我们研究中可采取的主要行动要点可总结如下：

- 在没有分布偏移的基线数据集上，所有方法都经过了良好校准，但在许多现实场景中却变得过度自信。因此，必须使用足够具有挑战性的基准测试来评估不确定性估计方法。否则，人们可能会误以为这些方法在现实世界部署期间会比实际情况更加可靠。

- 保形预测方法通常宣称具有理论覆盖保证，但这些保证依赖于一个在许多实际应用中不太可能成立的假设。因此，这些方法在现实场景中同样可能变得过度自信。如果从业者不对其潜在假设进行严格审视，这类理论保证可能会造成虚假的安全感，使这些模型更加不适合用于安全关键型部署。

- 对于基于特征空间密度的选择性预测方法，在合成数据集与真实世界数据集之间观察到的明显性能差异，是一个非常有趣的未来研究方向。如果能够理解造成这一性能间隙的原因，便有可能开发出一种能够在所有数据集上保持良好校准的不确定性估计方法。

- 基于特征空间密度的选择性预测方法相较于其他方法表现良好（鉴于其最先进的分布外检测性能，这一点符合预期），但在许多情况下也表现得过度自信。因此，仅比较不同方法的相对性能是不够的。

为了追踪我们是否正朝着真正可靠的不确定性估计方法这一最终目标取得实际进展，基准测试还必须从绝对意义上评估方法性能。

8 结论

我们提出了一个全面的基准测试，用于测试回归不确定性估计方法在真实世界分布偏移下的可靠性。该基准测试由 8 个公开可用的基于图像的回归数据集组成，涵盖了不同类型的、具有挑战性的分布偏移。我们使用该基准测试评估了许多最常见的不确定性估计方法，以及两种用于分布外检测的最先进不确定性分数。我们发现，尽管所有方法在不存在分布偏移时都经过了良好校准，但在许多基准数据集上，它们都变得过度自信。基于分布外不确定性分数的方法相较于其他方法表现良好，但其绝对性能仍远非完美。这揭示了当前回归不确定性估计方法的重要局限。因此，我们的工作向研究界提出了挑战：开发能够在所有基准数据集上实际生成校准预测区间的方法。为此，未来的研究方向包括探索在选择性预测中使用更复杂的模型来处理 $p(x)$——希望能够缩小合成数据集与真实世界数据集之间的性能差距——以及采用其他深度神经网络骨干架构。我们希望该基准测试能够推动更多研究，探索如何开发真正可靠的回归不确定性估计方法。

致谢

本研究得到了瑞典研究理事会通过项目 *NewLEADS – 动力系统学习的新方向* （合同编号：621-2016-06079）和深度概率回归 – 新模型与学习算法 （合同编号：2021-04301）的资助，并得到了 *Kjell & Märta Beijer*基金会的资助。我们还感谢 Ludvig Hult 和 Dave Zachariah 对早期稿件提出的宝贵意见。

参考文献

Michela Antonelli、Annika Reinke、Spyridon Bakas、Keyvan Farahani、Annette Kopp-Schneider、Bennett A Landman、Geert Litjens、Bjoern Menze、Olaf Ronneberger、Ronald M Summers 等。《医学分割十项全能挑战赛》。《自然·通讯》，13(1):1–13，2022年。

Spyridon Bakas、Hamed Akbari、Aristeidis Sotiras、Michel Bilello、Martin Rozycki、Justin S Kirby、John B Freymann、Keyvan Farahani 和 Christos Davatzikos。利用专家分割标签和影像组学特征，推进癌症基因组图谱胶质瘤磁共振成像数据集的发展。《科学数据》，4(1):1–13，2017年。

Spyridon Bakas、Mauricio Reyes、Andras Jakab、Stefan Bauer、Markus Rempfler、Alessandro Cri mi、Rus- sell TakeshiShinohara、Christoph Berger、Sung Min Ha、Martin Rozycki 等人。确定 BRATS挑战赛中脑肿瘤分割、进展评估和总生存期预测的最佳机器学习算法。*arXiv*预印本 *arXiv:1811.02629*，2018年。

Charles Blundell、Julien Cornebise、Koray Kavukcuoglu 和 Daan Wierstra。神经网络中的权重不确定性。载于国际机器学习会议（*ICML*），第 1613–1622 页，2015年。

Kurtland Chua、Roberto Calandra、Rowan McAllister 和 Sergey Levine。利用概率动力学模型，通过少量试验进行深度强化学习。载于神经信息处理系统进展（*NeurIPS*），第 4759–4770 页，2018年。

Youngseog Chung、Willie Neiswanger、Ian Char 和 Jeff Schneider。《超越弹珠损失：用于校准不确定性量化的分位数方法》。神经信息处理系统进展（*NeurIPS*），34：10971–10984，2021年。

James H Cole、Rudra PK Poudel、Dimosthenis Tsagkrasoulis、Matthan WA Caan、Claire Steves、Tim D Spector 和 Giovanni Montana。利用原始影像数据通过深度学习预测脑龄，可获得可靠且可遗传的生物标志物。*NeuroImage*，163：115–124，2017年。

Peng Cui、Wenbo Hu 和 Jun Zhu。利用最大均值差异实现校准可靠回归。神经信息处理系统进展 *(NeurIPS)*，33：17164–17175，2020年。Xin Ding、Yongwei Wang、Zuheng Xu、William J. Welch 和 Z. Jane Wang。用于图像生成的连续条件生成对抗网络：新型损失函数和标签输入机制。*arXiv* 预印本 *arXiv:2011.07466*，2020年。Xin Ding、Yongwei Wang、Zuheng Xu、William J Welch 和 Z. Jane Wang。CcGAN：用于图像生成的连续条件生成对抗网络。载于国际学习表征会议 *(ICLR)*，2021年。Samuel G Finlayson、Adarsh Subbaswamy、Karandeep Singh、John Bowers、Annabel Kupke、Jonathan Zittrain、Isaac S Kohane 和 Suchi Saria。人工智能中的临床医生与数据集偏移。《新英格兰医学杂志》，385(3)：283–286，2021年。Adam Fisch、Tommi S. Jaakkola 和 Regina Barzilay。校准选择性分类。机器学习研究汇刊 *(TMLR)*，2022年。Yarin Gal。深度学习中的不确定性。博士论文，剑桥大学，2016年。Yarin Gal 和 Zoubin Ghahramani。丢弃法作为贝叶斯近似：表示深度学习中的模型不确定性。载于国际机器学习会议 *(ICML)*，第1050–1059页，2016年。Yonatan Geifman 和 Ran El-Yaniv。深度神经网络的选择性分类。神经信息处理系统进展 *(NeurIPS)*，30，2017年。Tilmann Gneiting、Fadoua Balabdaoui 和 Adrian E Raftery。概率预测、校准与尖锐性。《英国皇家统计学会会刊：*B*辑（统计方法论）》，69(2)：243–268，2007年。Simon Graham、Quoc Dang Vu、Shan E Ahmed Raza、Ayesha Azam、Yee Wah Tsang、Jin Tae Kwak 和 Nasir Rajpoot。HoVer-Net：同时对多组织组织学图像中的细胞核进行分割和分类。《医学图像分析》，58：101563，2019年。Chuan Guo、Geoff Pleiss、Yu Sun 和 Kilian Q Weinberger。现代神经网络的校准问题。载于第*34*届国际机器学习会议 *(ICML)* 论文集，第1321–1330页，2017年。Fredrik K Gustafsson、Martin Danelljan、Goutam Bhat 和 Thomas B Schön。用于深度概率回归的基于能量的模型。载于欧洲计算机视觉会议 *(ECCV)* 论文集，第325–343页，2020a年。Fredrik K Gustafsson、Martin Danelljan 和 Thomas B Schön。评估用于鲁棒计算机视觉的可扩展贝叶斯深度学习方法。载于*IEEE/CVF*计算机视觉与模式识别会议 *(CVPR)* 研讨会论文集，2020b年。何恺明、Xiangyu Zhang、Ren Shaoqing 和 Sun Jian。用于图像识别的深度残差学习。载于*IEEE*计算机视觉与模式识别会议 *(CVPR)* 论文集，第770–778页，2016年。Dan Hendrycks 和 Thomas Dietterich。基准测试神经网络对常见损坏和扰动的鲁棒性。载于国际学习表征会议 *(ICLR)*，2019年。Dan Hendrycks、Steven Basart、Norman Mu、Saurav Kadavath、Frank Wang、Evan Dorundo、Rahul Desai、Tyler Zhu、Samyak Parajuli、Mike Guo 等。鲁棒性的多重面貌：对分布外泛化的批判性分析。载于*IEEE/CVF*国际计算机视觉会议 *(ICCV)* 论文集，第8340–8349页，2021a年。Dan Hendrycks、Kevin Zhao、Steven Basart、Jacob Steinhardt 和 Dawn Song。自然对抗样本。载于*IEEE/CVF*计算机视觉与模式识别会议 *(CVPR)* 论文集，第15262–15271页，2021b年。

Dan Hendrycks、Andy Zou、Mantas Mazeika、Leonard Tang、Bo Li、Dawn Song 和 Jacob Steinhardt。Pixmix：梦幻般的图像全面提升安全措施。载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，第16783–16792页，2022年。Radu Herbei 和 Marten H Wegkamp。带拒绝选项的分类。《加拿大统计学杂志》/《加拿大统计评论》，第709–721页，2006年。Eddy Ilg、Ozgun Cicek、Silvio Galesso、Aaron Klein、Osama Makansi、Frank Hutter 和 Thomas Bro。用于光流的不确定性估计与多假设网络。载于欧洲计算机视觉会议（*ECCV*）论文集，第652–667页，2018年。Pavel I zmailov、Patrick Nicholson、Sanae Lotfi 和 Andrew G Wilson。协变量偏移下贝叶斯模型平均的危险。神经信息处理系统进展（*NeurIPS*），34:3309–3322，2021年。Vilde Jensen、Filippo Maria Bianchi 和 Stian Norman Anfinsen。用于概率时间序列预测的集成保形化分位数回归。*arXiv* 预印本 *arXiv:2202.08756*，2022年。Wenming Jiang、Ying Zhao 和 Zehan Wang。用于回归深度神经网络模型的风险控制选择性预测。载于*2020*年国际神经网络联合会议（*IJCNN*），第1–8页。IEEE，2020年。Benedikt Atli Jónsson、Gyda Bjornsdottir、TE Thorgeirsson、Lotta María Ellingsen、G Bragi Walters、DF Gudbjartsson、Hreinn Stefansson、Kari Stefansson 和 MO Ulfarsson。利用深度学习进行脑龄预测，揭示相关序列变异。《自然·通讯》，10(1):5409，2019年。Alex Kendall 和 Yarin Gal。计算机视觉的贝叶斯深度学习需要哪些不确定性？载于神经信息处理系统进展（*NeurIPS*），第5574–5584页，2017年。Diederik P Kingma 和 Jimmy Ba。Adam：一种随机优化方法。*arXiv*预印本 *arXiv:1412.6980*，2014年。Roger Koenker 和 Gilbert Bassett Jr.。回归分位数。《计量经济学》：计量经济学会期刊，第33–50页，1978年。Pang Wei Koh、Shiori Sagawa、Henrik Marklund、Sang Michael Xie、Marvin Zhang、Akshay Balsubramani、Weihua Hu、Michihiro Yasunaga、Richard Lanas Phillips、Irena Gao 等。WILDS：真实环境中分布偏移的基准测试。载于国际机器学习会议（*ICML*），第5637–5664页。PMLR，2021年。Johnson Kuan 和 Jonas Mueller。回归基础：重新审视分布外检测基线。*arXiv*预印本 *arXiv:2207.03061*，2022年。Volodymyr Kuleshov、Nathan Fenner 和 Stefano Ermon。使用校准回归实现深度学习的准确不确定性估计。载于国际机器学习会议（*ICML*），第2796–2804页。PMLR，2018年。Neeraj Kumar、Ruchika Verma、Sanuj Sharma、Surabhi Bhargava、Abhishek Vahadane 和 Amit Sethi。用于计算病理学的广义细胞核分割数据集与技术。*IEEE*医学影像汇刊，36(7):1550–1560，2017年。Balaji Lakshminarayanan、Alexander Pritzel 和 Charles Blundell。利用深度集成进行简单且可扩展的预测不确定性估计。载于神经信息处理系统进展（*NeurIPS*），第6402–6413页，2017年。Taro Langner、Robin Strand、Håkan Ahlström 和 Joel Kullberg。基于英国生物样本库全身磁共振成像数据、采用可解释神经网络回归的大规模生物测量学。《科学报告》，10(1):17752，2020年。Taro Langner、Fredrik K Gustafsson、Benny Avelin、Robin Strand、Håkan Ahlström 和 Joel Kullberg。利用英国生物样本库磁共振成像数据和深度回归集成进行不确定性感知的身体成分分析。计算机化医学影像与图形学，93:101994，2021年。

Stéphane Lathuilière、Pablo Mesejo、Xavier Alameda-Pineda 和 Radu Horaud。深度回归的全面分析。*IEEE*模式分析与机器智能汇刊（*TPAMI*），2019年。

Hei Law 和 Jia Deng。Cornernet：将物体检测为成对关键点。载于《欧洲计算机视觉会议（*ECCV*）论文集，第734–750页，2018年。Dan Levi、Liran Gispan、Niv Giladi 和 Ethan Fetaya。评估并校准回归任务中的不确定性预测。*Sensors*，22(15):5540，2022年。Yi-An Ma、Tianqi Chen 和 Emily Fox。随机梯度马尔可夫链蒙特卡罗的完整方案。载于《神经信息处理系统进展（*NeurIPS*）》，第2917–2925页，2015年。Emmanuel Maggiori、Yuliya Tarabalka、Guillaume Charpiat 和 Pierre Alliez。语义标注方法能否推广到任何城市？Inria航拍图像标注基准。载于*2017 IEEE*国际地球科学与遥感研讨会（*IGARSS*），第3226–3229页。IEEE，2017年。Andrey Malinin、Neil Band、Yarin Gal、Mark Gales、Alexander Ganshin、German Chesnokov、Alexey Noskov、Andrey Ploskonosov、Liudmila Prokhorenkova、Ivan Provilkov、Vatsal Raina、Vyas Raina、Denis Roginskiy、Mariya Shmatova、Panagiotis Tigas 和 Boris Yangel。Shifts：一个涵盖多个大规模任务中真实分布偏移的数据集。载于第三十五届神经信息处理系统大会（*NeurIPS*），数据集与基准测试专题，2021年。Bjoern H Menze、Andras Jakab、Stefan Bauer、Jayashree Kalpathy-Cramer、Keyvan Farahani、Justin Kirby、Yuliya Burren、Nicole Porz、Johannes Slotboom、Roland Wiest 等。多模态脑肿瘤图像分割基准（BRATS）。*IEEE*医学影像汇刊，34(10):1993–2024，2014年。Takeru Miyato、Toshiki Kataoka、Masanori Koyama 和 Yuichi Yoshida。用于生成对抗网络的谱归一化。载于国际学习表征会议（*ICLR*），2018年。

Jishnu Mukhoti、Andreas Kirsch、Joost van Amersfoort、Philip HS Torr 和 Yarin Gal。深度确定性不确定性：一种简单的基线。*arXiv* 预印本，2021年a。

Jishnu Mukhoti、Andreas Kirsch、Joost van Amersfoort、Philip HS Torr 和 Yarin Gal。具有适当归纳偏置的确定性神经网络能够捕捉认知不确定性和偶然不确定性。*arXiv* 预印本 *arXiv:2102.11582*，2021年b。

Peter Naylor、Marick Laé、Fabien Reyal 和 Thomas Walter。通过对距离图进行深度回归，实现组织病理图像中的细胞核分割。*IEEE*医学影像汇刊，38(2):448–459，2018年。

Radford M Neal。神经网络的贝叶斯学习。博士论文，多伦多大学，1995年。

David Ouyang、Bryan He、Amirata Ghorbani、Neal Yuan、Joseph Ebinger、Curtis P Langlotz、Paul A Heidenreich、Robert A Harrington、David H Liang、Euan A Ashley 等。用于逐搏评估心脏功能的视频人工智能。《自然》，580(7802):252–256，2020年。

Yaniv Ovadia、Emily Fertig、Jie Ren、Zachary Nado、D. Sculley、Sebastian Nowozin、Joshua Dillon、Balaji Lakshminarayanan 和 Jasper Snoek。你能相信模型的不确定性吗？在数据集偏移下评估预测不确定性。载于神经信息处理系统进展（*NeurIPS*），第 32 卷，2019年。

Harris Papadopoulos。归纳保形预测：理论及其在神经网络中的应用。载于人工智能工具。Citeseer，2008年。

Harris Papadopoulos、Kostas Proedrou、Volodya Vovk 和 Alex Gammerman。用于回归的归纳置信机。载于欧洲机器学习会议，第 345–356 页。Springer，2002年。

Adam Paszke、Sam Gross、Francisco Massa、Adam Lerer、James Bradbury、Gregory Chanan、Trevor Killeen、Zeming Lin、Natalia Gimelshein、Luca Antiga 等。《PyTorch：一种命令式风格的高性能深度学习库》。发表于神经信息处理系统进展（*NeurIPS*），第 8024–8035 页，2019年。

Fabian Pedregosa、Gaël Varoquaux、Alexandre Gramfort、Vincent Michel、Bertrand Thirion、Olivier Grisel、Mathieu Blondel、Peter Prettenhofer、Ron Weiss、Vincent Dubourg、Jake Vanderplas、Alexandre Passos、David Cournapeau、Matthieu Brucher、Matthieu Perrot 和 Édouard Duchesnay。《Scikit-learn：Python 中的机器学习》。机器学习研究期刊（*JMLR*），12(85):2825–2830，2011年。

Ethan Pickering 和 Themistoklis P Sapsis。《用于量化不确定性质量的结构与分布度量：评估用于回归的高斯过程、深度神经网络和深度神经算子》。*arXiv* 预印本 *arXiv:2203.04515*，2022年。

Geoff Pleiss、Amauri Souza、Joseph Kim、Boyi Li 和 Kilian Q. Weinberger。用于回归任务的神经网络分布外检测，2020年。网址：`https://openreview.net/forum?id=ryxsUySFwr`。

Ryan Poplin、Avinash V Varadarajan、Katy Blumer、Yun Liu、Michael V McConnell、Greg S Corrado、Lily Peng 和 Dale R Webster。通过深度学习从视网膜眼底照片预测心血管风险因素。《自然·生物医学工程》，2(3):158–164，2018年。

Joaquin Quionero-Candela、Masashi Sugiyama、Anton Schwaighofer 和 Neil D Lawrence。机器学习中的数据集偏移，2009年。

Yaniv Romano、Evan Patterson 和 Emmanuel Candes。保形化分位数回归。神经信息处理系统进展（*NeurIPS*），32，2019年。

Rasmus Rothe、Radu Timofte 和 Luc Van Gool。无需面部标志点、从单张图像预测真实年龄与表观年龄的深度期望。《国际计算机视觉期刊》，126(2-4):144–157，2016年。

Mohammadreza Salehi、Hossein Mirzaei、Dan Hendrycks、Yixuan Li、Mohammad Hossein Rohban 和 Mohammad Sabokrou。关于异常、新颖性、开放集及分布外检测的综合综述：解决方案与未来挑战。机器学习研究汇刊（*TMLR*），2022年。

Leo Schwinn、Leon Bungert、An Nguyen、René Raab、Falk Pulsmeyer、Doina Precup、Björn Eskofier 和 Dario Zanca。通过决策区域量化提升对真实世界和最坏情况分布偏移的鲁棒性。发表于国际机器学习会议（*ICML*），第 19434–19449 页。PMLR，2022年。

Joan Serrà、David Álvarez、Vicenç Gómez、Olga Slizovskaia、José F. Núñez 和 Jordi Luque。《基于似然的生成模型中的输入复杂度与分布外检测》。发表于国际学习表征会议（*ICLR*），2020年。

Shaoshuai Shi、Xiaogang Wang 和 Hongsheng Li。《PointRCNN：基于点云的三维目标提议生成与检测》。发表于*IEEE*计算机视觉与模式识别会议论文集（*CVPR*），第770–779页，2019年。

Wen Shi、Guohui Yan、Yamin Li、Haotian Li、Tingting Liu、Cong Sun、Guangbin Wang、Yi Zhang、Yu Zou 和 Dan Wu。《利用基于注意力的深度集成与不确定性进行胎儿脑龄估计和异常检测》。*NeuroImage*，223:117316，2020年。

Amber L Simpson、Michela Antonelli、Spyridon Bakas、Michel Bilello、Keyvan Farahani、Bram Van Ginneken、Annette Kopp-Schneider、Bennett A Landman、Geert Litjens、Bjoern Menze 等人。《用于开发和评估分割算法的大型标注医学图像数据集》。*arXiv* 预印本 *arXiv:1902.09063*，2019年。Ingo Steinwart 和 Andreas Christmann。《借助弹珠损失估计条件分位数》。*Bernoulli*，17(1):211–225，2011年。Yiyou Sun、Yifei Ming、Xiaojin Zhu 和 Yixuan Li。《利用深度最近邻进行分布外检测》。发表于国际机器学习会议（*ICML*），2022年。Eric J Topol。《高性能医学：人类智能与人工智能的融合》。*NatureMedicine*，25(1):44–56，2019年。

Philipp Tschandl、Cliff Rosendahl 和 Harald Kittler。《HAM10000 数据集：一个包含常见色素性皮肤病变多源皮肤镜图像的大型集合》。《科学数据》，5(1):1–9，2018年。

Florian Wenzel、Andrea Dittadi、Peter Gehler、Carl-Johann Simon-Gabriel、Max Horn、Dominik Zietlow、David Kernert、Chris Russell、Thomas Brox、Bernt Schiele 等。检验迁移学习中的分布外泛化。《神经信息处理系统进展（*NeurIPS*）》，35：7181–7198，2022年。

OliviaWiles、SvenGowal、Florian Stimberg、Sylvestre-AlviseRebuffi、IraKtena、Krishnamurthy Dj Dvijotham 和 Ali Taylan Cemgil。对分布偏移进行细粒度分析。载于国际学习表征会议（*ICLR*），2022年。

Bin Xiao、Haiping Wu 和 Yichen Wei。用于人体姿态估计和跟踪的简单基线。载于欧洲计算机视觉会议（*ECCV*）论文集，第466–481页，2018年。

Wufeng Xue、Ali Islam、Mousumi Bhaduri 和 Shuo Li。通过联合表征与回归学习直接估计多种心脏指标。*IEEE*医学影像汇刊，36(10):2057–2067，2017年。

Huaxiu Yao、Yu Wang、Sai Li、Linjun Zhang、Weixin Liang、James Zou 和 Chelsea Finn。通过选择性增强提升分布外鲁棒性。载于国际机器学习会议 *(ICML)*，第 25407–25437 页。PMLR，2022年。

Christopher Yeh、Anthony Perez、Anne Driscoll、George Azzari、Zhongyi Tang、David Lobell、Stefano Ermon 和 Marshall Burke。利用公开可用的卫星图像和深度学习了解非洲的经济福祉。《自然·通讯》，11(1):1–11，2020年。

Sheheryar Zaidi、Arber Zela、Thomas Elsken、Chris C Holmes、Frank Hutter 和 Yee Teh。用于不确定性估计和数据集偏移的神经集成搜索。神经信息处理系统进展 *(NeurIPS)*，34:7898–7911，2021年。

Marvin Mengxin Zhang、Sergey Levine 和 Chelsea Finn。MEMO：通过适应和增强提升测试时鲁棒性。发表于神经信息处理系统进展（*NeurIPS*），2022年。Zheng Zhu、Qiang Wang、Bo Li、Wei Wu、Junjie Yan 和 Weiming Hu。用于视觉目标跟踪的干扰物感知孪生网络。发表于欧洲计算机视觉会议（*ECCV*）论文集，第101–117页，2018年。

数据集详情

下面将详细介绍第3.1节中的12个数据集。

细胞 给定一张合成荧光显微图像 x，任务是预测图像中的细胞数量 y。我们使用 Ding 等人（2021年；2020年）的 Cell-200 数据集，其中包含 200 000 张大小为 64×64 的灰度图像。我们随机抽取 10 000 张训练图像、2 000 张验证图像和 10 000 张测试图像。因此，训练/验证集与测试集之间不存在分布偏移。因此，我们将其用作基线数据集。

细胞-尾部 我们创建了一个 Cells 的变体，其中训练/验证集与测试集之间存在明显的分布偏移。对于 10 000 张训练图像和 2 000 张验证图像，回归目标 y 被限制在范围 $]50, 150]$内。对于 10 000 张测试图像，目标则位于完整的原始范围 $[1, 200]$内。

Cells-Gap 我们创建了另一个 Cells 变体，其中训练/验证集与测试集之间存在明显的分布偏移。对于 10 000 张训练图像和 2 000 张验证图像，回归目标 y 被限制在 $[1, 50[\cup]150, 200]$内。对于 10 000 张测试图像，目标则位于完整的原始范围 $[1, 200]$内。

椅子角度 给定一张 x 渲染椅子模型的合成图像，任务是预测椅子的偏航角y。我们使用 Ding 等人（2021年；2020年）的 RC-49 数据集，其中包含 64×64 张不同椅子模型的 RGB 图像，这些模型以 899 个偏航角进行渲染，角度范围从 $0.1°$ 到 $89.9°$，步长为 $0.1°$。我们随机划分其训练集，得到 17 640 张训练图像和 4 410 张验证图像。通过对其测试集进行子采样，我们还得到 11 225 张测试图像。训练/验证集与测试集之间没有明显的分布偏移，因此我们将其用作第二个基线数据集。

ChairAngle-Tails 我们创建了 ChairAngle 的一个变体，使训练/验证集与测试集之间存在明显的分布偏移。对于训练集和验证集，我们将回归目标 y 限制在 $]15, 75[$范围内。对于测试集，目标则位于原始完整范围 $]0, 90[$内。我们得到 11 760 张训练图像、2 940 张验证图像和 11 225 张测试图像。

ChairAngle-Gap ChairAngle 的另一个变体，在训练/验证集与测试集之间存在明显的分布偏移。对于 11 760 张训练图像和 2 940 张验证图像，回归目标 y 被限制在 $]0,\ 30[\cup]60,\ 90[$。对于 11 225 张测试图像，目标则位于完整的原始范围 $]0, 90[$内。

资产财富 给定一张卫星图像 x（包含 8 个图像通道），任务是预测该区域的资产财富指数 y。我们使用 PovertyMap-Wilds 数据集（Koh 等，2021），这是 Yeh 等人（2020年）收集的数据集的一个变体。我们使用数据中的训练集、验证集-ID 和测试集-OOD 子集，共包含 9 797 张训练图像、1 000 张验证图像和 3 963 张测试图像。我们将图像尺寸从 224×224 调整为 64×64。训练/验证集和测试集包含来自互不重叠的非洲国家集合的卫星图像，从而形成分布偏移。

心室容积 给定一张人类心脏的超声心动图图像 x，任务是预测左心室的容积y。我们使用 Ouyang 等人（2020年）提供的 EchoNet-Dynamic 数据集，其中包含 10 030 段超声心动图视频。每段视频记录一个完整的心动周期，并标注了两个不同时间点的左心室容积测量值，分别代表收缩末期（收缩结束时，容积较小）和舒张末期（收缩开始前，容积较大）。对于每段视频，我们仅提取其中一个容积测量值及其对应的视频帧。为了在训练/验证集和测试集之间形成明显的分布偏移，我们在训练集和验证集中选取收缩末期容积（较小的容积），而在测试集中选取舒张末期容积（较大的容积）。我们采用数据集提供的划分，得到 7 460 张训练图像、1 288 张验证图像和 1 276 张测试图像。我们将图像尺寸从 112×112 调整为 64×64。

脑肿瘤像素 给定一张脑部 MRI 扫描的图像切片 x，任务是预测图像中被标记为脑肿瘤的像素数量 y。我们使用医学分割十项全能挑战赛的脑肿瘤数据集（Simpson 等人，2019年；Antonelli 等人，2022年），该数据集是2016年和2017年 BraTS 挑战赛所使用数据的子集（Bakas 等人，2018年；2017年；Menze 等人，2014年）。该数据集包含484次脑部 MRI 扫描及其对应的肿瘤分割掩膜。我们按80%/20%/20%的比例将这些扫描划分为训练集、验证集和测试集。这些扫描是大小为 $240 \times 240 \times 155$ 的三维体积。我们将每次扫描转换为155张大小为 240×240 的图像切片，并通过统计每张图像中被标记的脑肿瘤像素数量，为每张图像创建一个回归目标。随后，我们还移除了所有不包含肿瘤像素的图像。原始图像切片包含4个通道（液体衰减反转恢复、T1 加权、T1 增强、T2 加权），但我们只使用前3个通道，并将切片转换为标准 RGB 图像。最终得到20 614张训练图像、6 116张验证图像和6 252张测试图像。我们还将图像从大小 240×240 调整为 64×64。

皮肤病变像素 给定一张色素性皮肤病变的皮肤镜图像 x，任务是预测图像中被标记为病变的像素数量 y。我们使用 Tschandl 等人（2018年）提供的 HAM10000 数据集，该数据集包含10 015张皮肤镜图像及其对应的皮肤病变分割掩膜。HAM10000由四个不同的子数据集组成，其中三个（ViDIR Legacy、ViDIR Current 和 ViDIR MoleMax）采集于奥地利，第四个子数据集（Rosendahl）采集于澳大利亚。为了在训练集/验证集与测试集之间形成明确的分布偏移，我们使用澳大利亚子数据集（Rosendahl）作为测试集。将剩余图像随机按85%/15%的比例划分为训练集和验证集后，我们得到6 592张训练图像、1 164张验证图像和2 259张测试图像。随后，我们通过统计每张图像中被标记的皮肤病变像素数量，为每张图像创建一个回归目标。我们还将图像从大小 450×600 调整为 64×64。

HistologyNucleiPixels 给定一张苏木精-伊红染色组织学图像 x，任务是预测图像中被标记为细胞核的像素数量 y。我们使用 Graham 等人（2019年）的 CoNSeP 数据集，以及他们提供的 Kumar 数据集（Kumar 等人，2017年）和 TNBC 数据集（Naylor 等人，2018年）的预处理版本。这些数据集包含大幅苏木精-伊红染色图像块（尺寸为 $1\,000 \times 1\,000$ 或 512×512），并以 $40\times$ 物镜放大倍数拍摄，同时提供对应的细胞核分割掩膜。这三个数据集采集自不同的医院/机构，标本保存和染色流程也各不相同。因此，使用 CoNSeP 和 Kumar 作为训练/验证集，并使用 TNBC 作为测试集，可以形成明显的分布偏移。我们从大幅图像块中通过规则网格划分提取 64×64 图像切片，并通过计算每个图像切片中被标记的细胞核像素数量，为其创建回归目标。随后，我们还移除了所有不含细胞核像素的图像。最终，我们得到 10 808 张训练图像、2 702 张验证图像和 2 267 张测试图像。

AerialBuildingPixels 给定一张航拍图像 x，任务是预测图像中被标记为建筑物的像素数量 y。我们使用 Inria 航拍图像标注数据集（Maggiori 等人，2017年），其中包含 180 张大型航拍图像及相应的建筑物分割掩码。这些图像的尺寸为$5\,000 \times 5\,000$，拍摄于五个不同的地理位置。我们使用来自两座人口稠密的美国城市（奥斯汀和芝加哥）的图像作为训练/验证集，并使用来自欧洲较偏远地区（奥地利西蒂罗尔）的图像作为测试集，从而形成明显的分布偏移。我们首先将图像缩放至尺寸 $1\,000 \times 1\,000$，然后通过规则网格划分提取 64×64 图像切片。我们还通过统计每个图像切片中被标记为建筑物的像素数量，为其创建一个回归目标。移除所有不含建筑物像素的图像后，最终得到 11 184 张训练图像、2 797 张验证图像和 3 890 张测试图像。

图 4 中 Cells 和 ChairAngle 的其他变体规定如下。Cells：训练/验证集与测试集之间的回归目标范围没有差异。细胞-尾部：训练/验证集的目标范围为 $]50, 150]$，测试集的目标范围为 $[1, 200]$。我们为测试集创建三个具有中间目标范围的版本（1：$[37.5, 163.5]$，2：$[25, 176]$, 3：$[12.5, 188.5]$）。ChairAngle：训练/验证集与测试集之间的目标范围没有差异。ChairAngle-Gap：训练/验证集的目标范围为 $]0,\ 30[\cup]60,\ 90[$，测试集的目标范围为 $]0, 90[$。我们为测试集创建三个具有中间目标范围的版本（1：$]0,\ 33.75[\cup]56.25,\ 90[$, 2：$]0,\ 37.5[\cup]52.5,\ 90[$, 3：$]0,\ 41.25[\cup]48.75,\ 90[$）。

B 附加结果与方法变体

请注意，表 *A1* 至表 *A42* 中的结果并非仅四舍五入/截断至有效数字。

为确保测试集覆盖率结果不受我们选择研究 90% 预测区间 ($\alpha = 0.1$) 的影响，我们针对另外两种失覆盖率 α 重复了大部分评估。具体而言，我们在 9/12 个数据集上，使用 80% ($\alpha = 0.2$) 和 95% ($\alpha = 0.05$) 预测区间重新评估了 6/10 种方法。80% 预测区间的结果见图 A7 至图 A10 和表 A13 至表 A21；95% 预测区间的结果见图 A11 至图 A14 和表 A22 至表 A30。总体而言，我们观察到的趋势非常相似。例如，在细胞和椅子角度数据集上，所有方法的预测区间仍几乎完全校准，即它们得到的测试集覆盖率都非常接近 80%/95%；但在细胞-尾部和 ChairAngle-Gap 上，只有高斯 + 选择性高斯混合模型仍保持良好校准。除 BrainTumourPixels 外，所有方法在所有真实世界数据集上也都显著过度自信。

图 A15 展示了五种常见回归不确定性估计方法在 AerialBuildingPixels 数据集及其两个测试集不同的版本上的测试集覆盖率结果。对于这三个数据集，训练/验证集均包含来自奥斯汀和芝加哥的图像。对于 AerialBuildingPixels，测试集包含来自

奥地利西蒂罗尔的图像。对于 AerialBuildingPixels-Kitsap，测试集则包含来自华盛顿州基萨普县的图像。对于 AerialBuildingPixels-Vienna，测试集包含来自奥地利维也纳的图像。直观而言，AerialBuildingPixels-Kitsap 和 AerialBuildingPixels-Vienna 的训练/验证集与测试集之间的分布偏移可能小于原始 AerialBuildingPixels，但我们在图 A15 中并未观察到明显趋势。

图 A16 和 A17 展示了一项研究，我们旨在将测试集覆盖率性能与分布偏移的定量度量（训练/验证集与测试集分布之间的“距离”）联系起来，以补充我们在第 7 节真实世界数据集之间的性能差异大多合乎逻辑段落中的定性讨论。然而，如何量化真实世界数据集中的分布偏移程度并不显而易见，参见例如 Wenzel 等人（2022）的附录 E.1。我们在此探究，验证集和测试集上的回归准确度差异（平均绝对误差）能否用作此类度量，并将 Wenzel 等人（2022）的方法扩展到我们的回归场景。我们针对五种常见的回归不确定性估计方法计算平均测试覆盖误差，而验证集/测试集的平均绝对误差则针对保形预测方法（标准直接回归模型）计算。六个合成数据集的结果见图 A16，六个真实世界数据集的结果见图 A17。我们观察到，总体而言，较大的分布偏移度量似乎确实对应着较差的测试集覆盖率性能。在这 12 个数据集中，HistologyNucleiPixels 是唯一一个明显偏离这一总体趋势的数据集。

除了第 4 节中规定的 10 种方法的主要对比外，我们还评估了几种方法变体。这些实验的结果见表 A31 至表 A42。对于高斯 + 选择性高斯混合模型，我们将高斯混合模型混合成分的数量从标准的 $k = 4$ 调整为 $k = 2$和 $k = 8$，但未观察到结果中存在特别一致或显著的趋势。同样，对于高斯 + 选择性 k 近邻，我们将近邻数量从标准的 $k = 10$ 调整为 $k = 5$ 和 $k = 20$，但同样未发现明显趋势。对于高斯 + 选择性 k 近邻，我们还尝试将余弦相似度距离度量替换为 L2 距离，得到的结果仍然非常相似。最后，遵循 Mukhoti 等人（2021年）的做法，我们加入谱归一化（Miyato 等人，2018年）以进一步进行特征空间正则化，但对于高斯 + 选择性高斯混合模型，并未观察到显著改进。

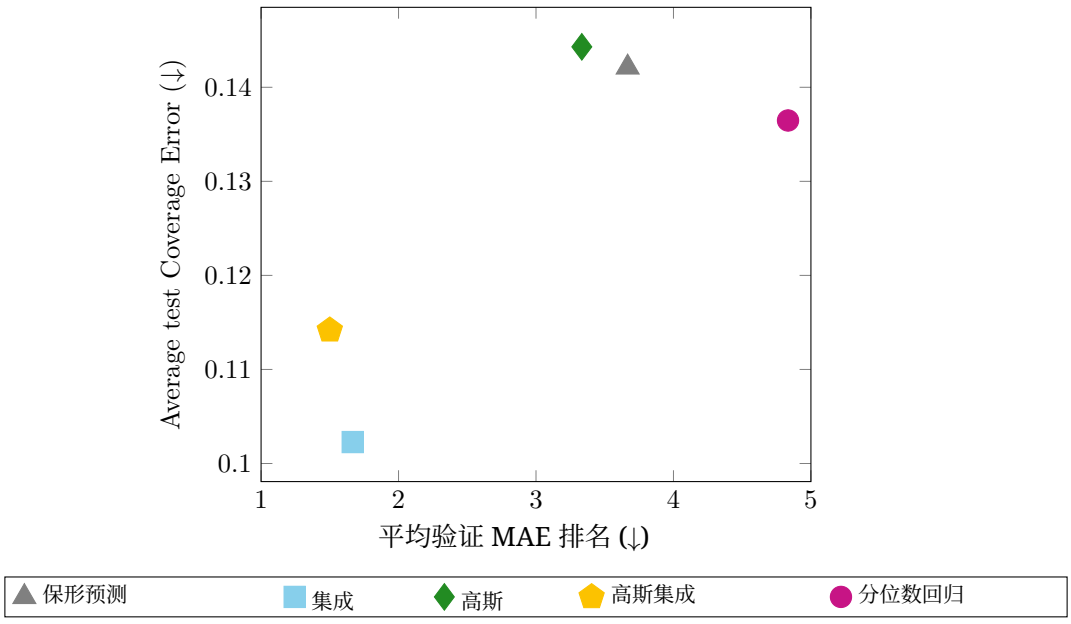


图 A1：五种常见回归不确定性估计方法在六个真实世界数据集上的性能比较，比较指标为平均测试覆盖误差和平均验证 MAE 排名（在每个数据集上，根据验证集平均绝对误差对这五种方法进行 1–5 名排名，然后计算平均排名）。集成和高斯集成表现最佳。

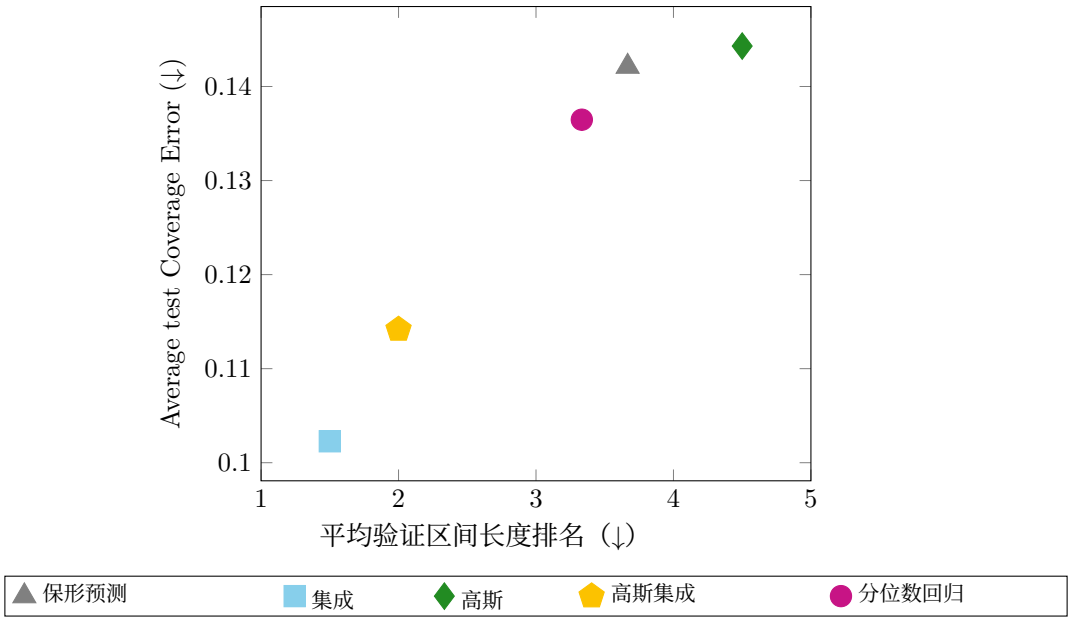
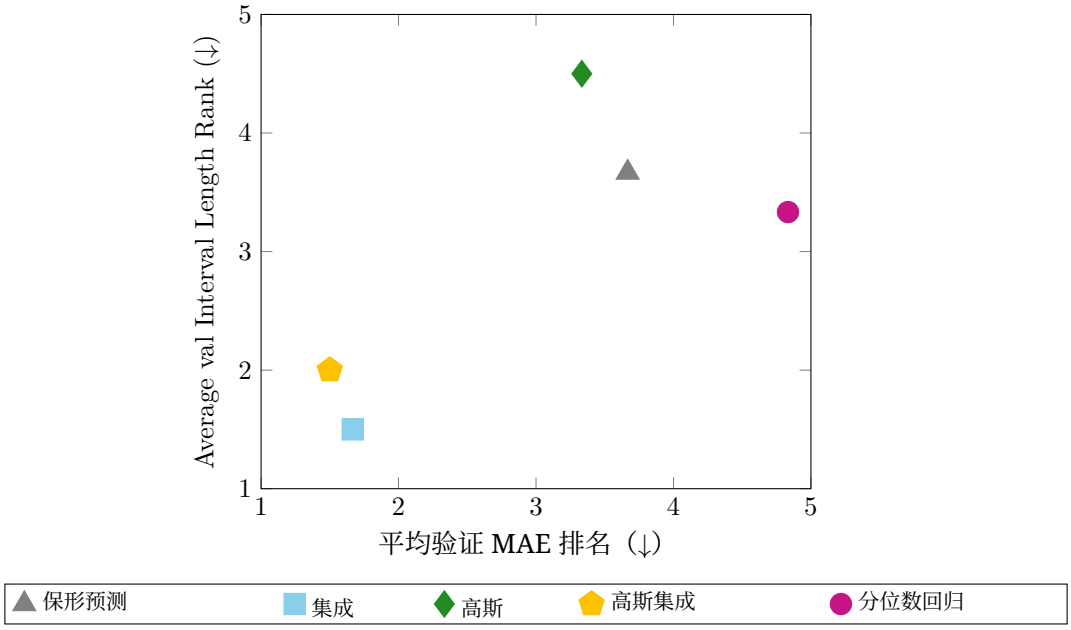


图 A2：在六个真实世界数据集上，对五种常见回归不确定性估计方法的性能进行比较，比较指标为平均测试覆盖误差和平均验证集区间长度排名（五种方法分别根据其在每个数据集上的验证集区间长度排名为 1～5，然后计算平均排名）。集成和高斯集成的表现最佳。



图A3：五种常见回归不确定性估计方法在六个真实世界数据集上的性能比较，比较指标为平均验证区间长度排名和平均验证 MAE 排名。集成和高斯集成表现最佳。

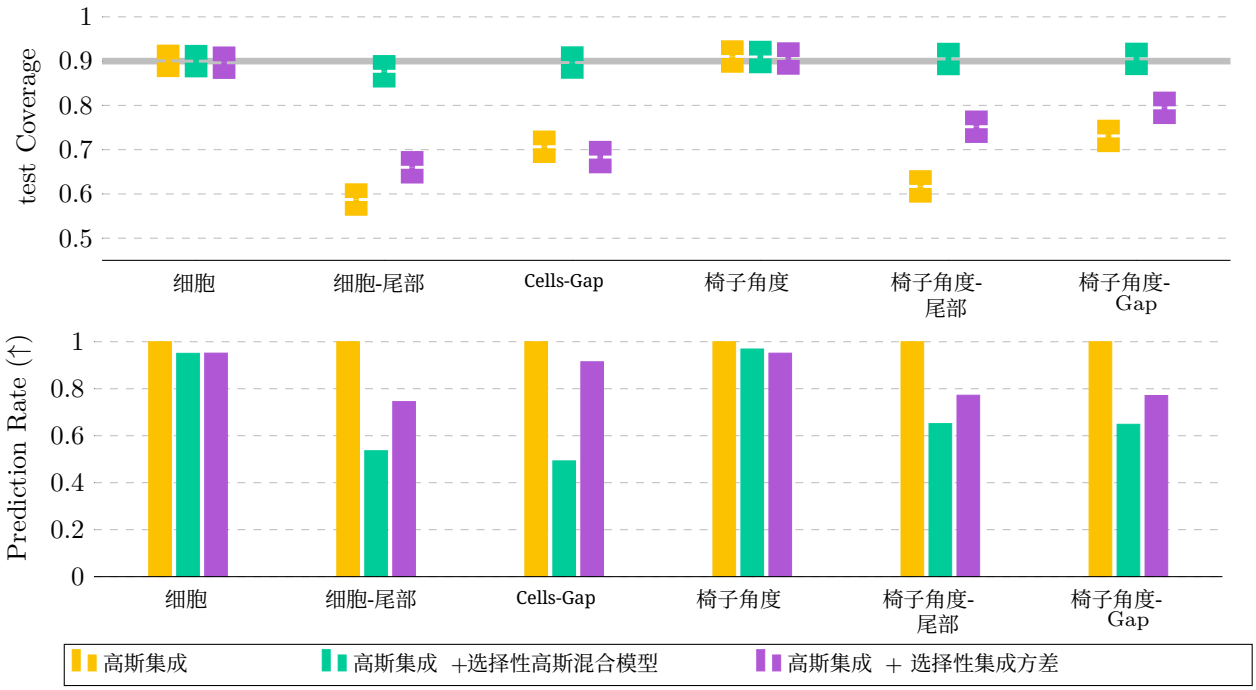


图 A4：基于高斯集成的两种选择性预测方法在六个合成 数据集上的结果。**上：** 测试集覆盖率结果。**下：** 测试集预测率结果。

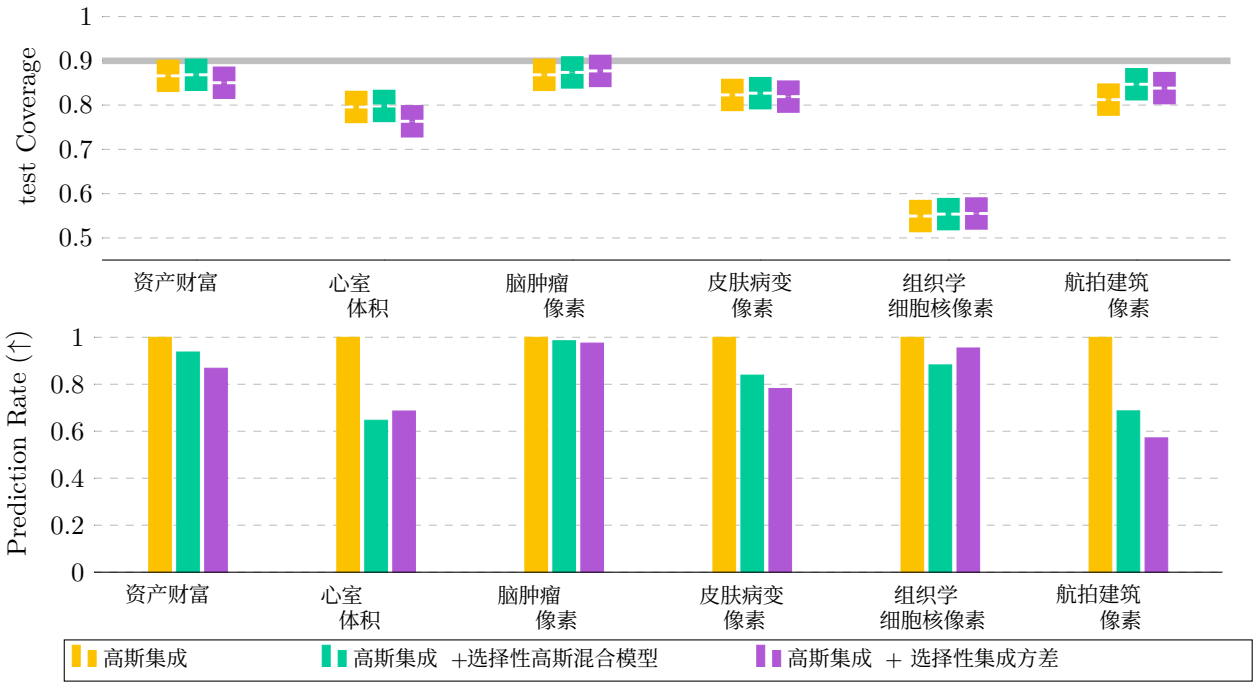


图 A5：基于高斯集成的两种选择性预测方法在真实世界数据集上的结果。**上：** 测试集覆盖率方面的结果。**下：** 测试集预测率方面的结果。

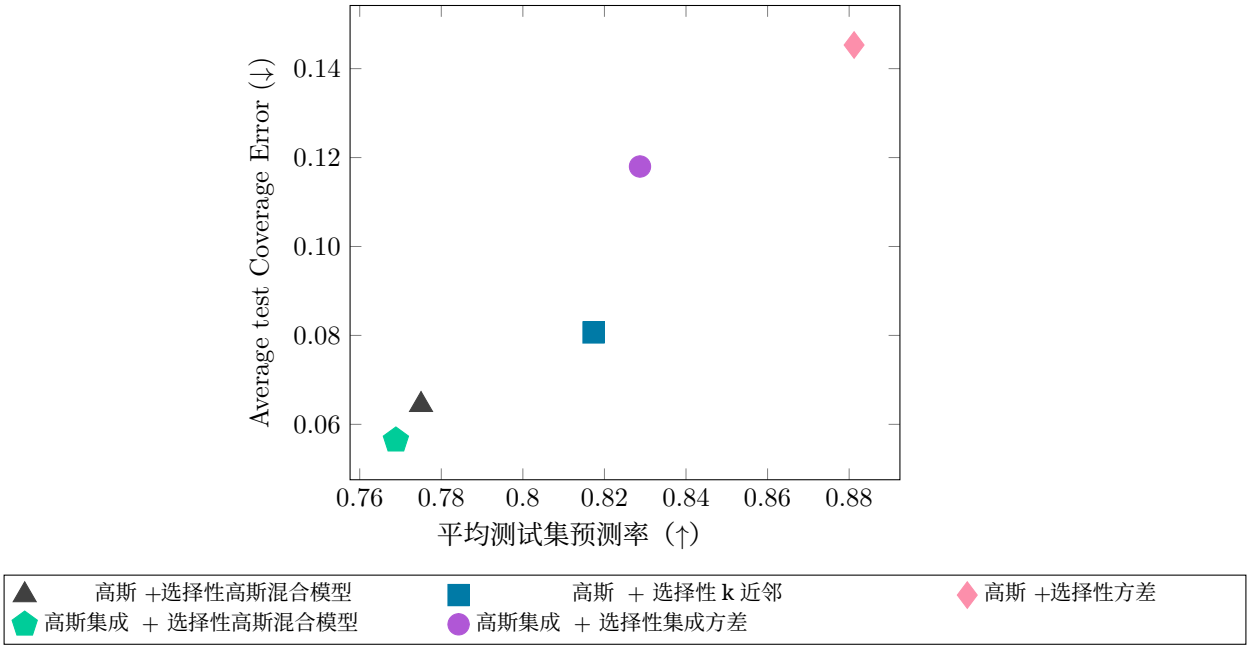


图 A6：在平均测试覆盖误差和测试预测率方面，对 12 个数据集上的选择性预测方法进行性能比较。高斯集成 + 选择性高斯混合模型的覆盖误差最低，但预测率也最低。对于这五种方法，覆盖误差每降低一次，预测率也会相应下降。

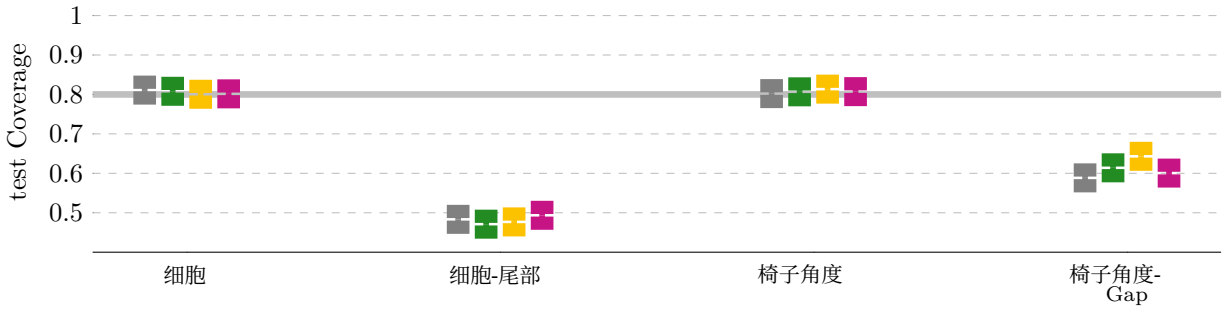


图 A7：失覆盖率 $\alpha = 0.2$：在四个合成数据集上，针对四种常见的回归不确定性估计方法（保形预测、高斯、高斯集成、分位数回归）的测试集覆盖率结果。其他指标请参见表 A13–表 A16。

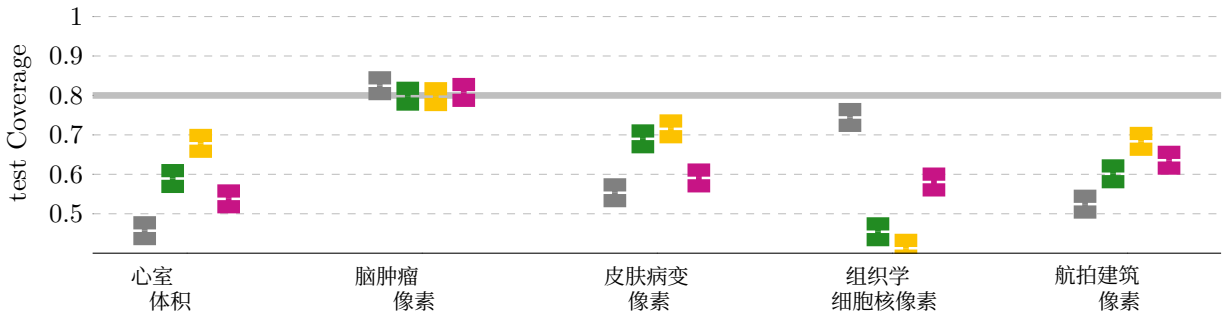


图 A8：失覆盖率 $\alpha = 0.2$：四种常见回归不确定性估计方法（保形预测、高斯、高斯集成、分位数回归）在五个真实世界 数据集上的测试集覆盖率结果。其他指标请参见表 A17–表 A21。

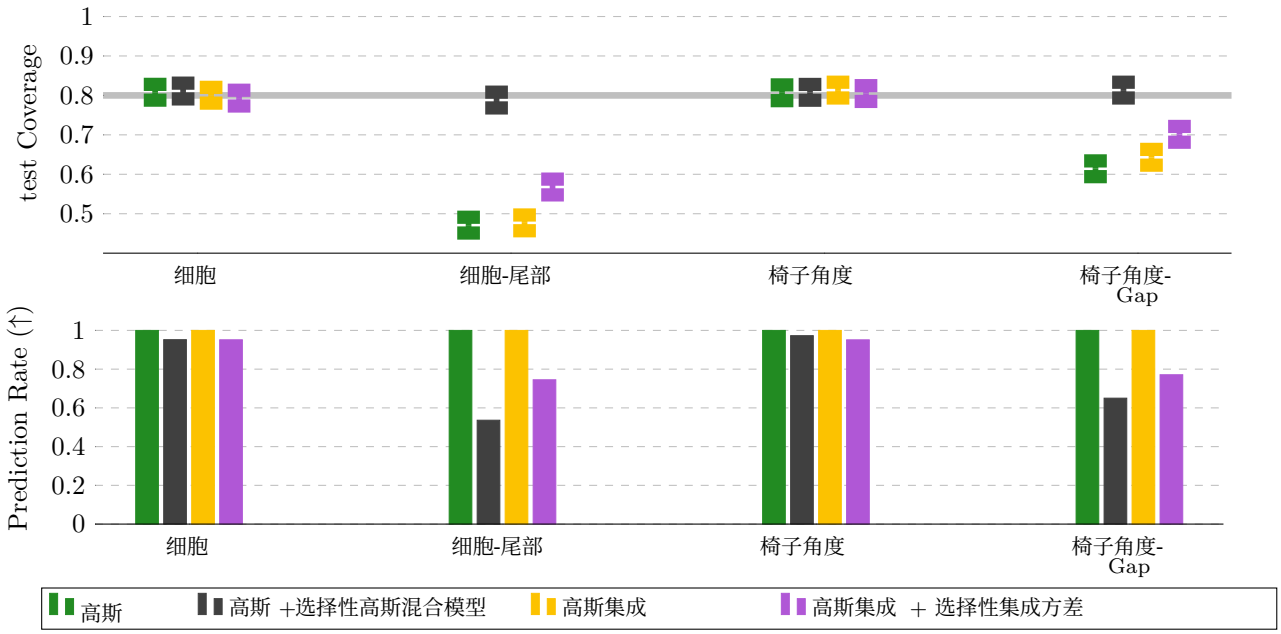


图 A9：失覆盖率 $\alpha = 0.2$：在四个合成数据集上，针对两种选择性预测方法的结果。其他指标请参见表 A13—表 A16。

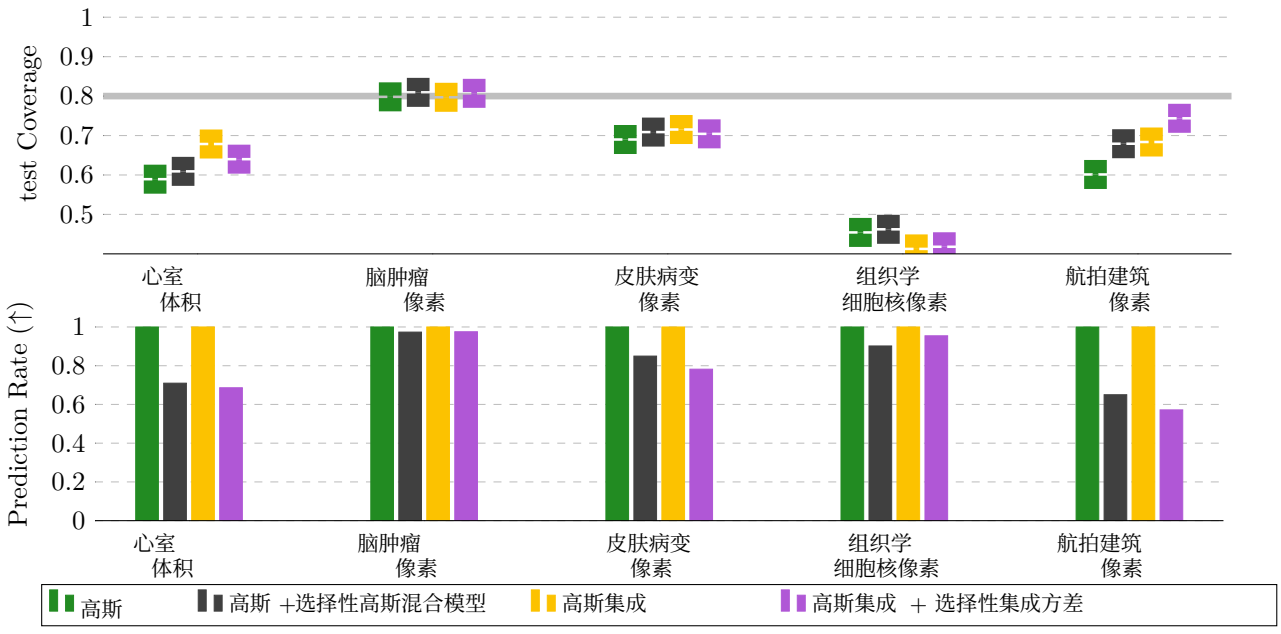


图 A10：失覆盖率 $\alpha = 0.2$：两种选择性预测方法在五个真实世界数据集上的结果。其他指标请参见表 A17 - 表 A21。

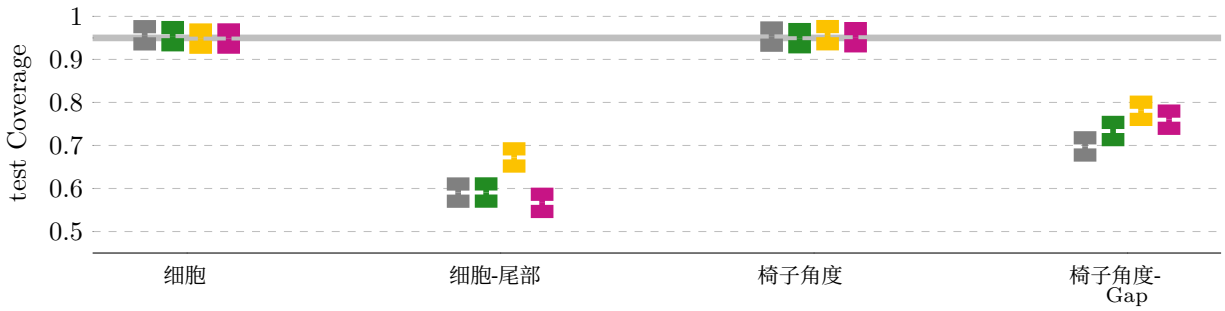


图 A11：失覆盖率 $\alpha = 0.05$：四种常见回归不确定性估计方法（保形预测、高斯、高斯集成、分位数回归）在四个合成 数据集上的测试集覆盖率结果。其他指标请参见表 A22 - 表 A25。

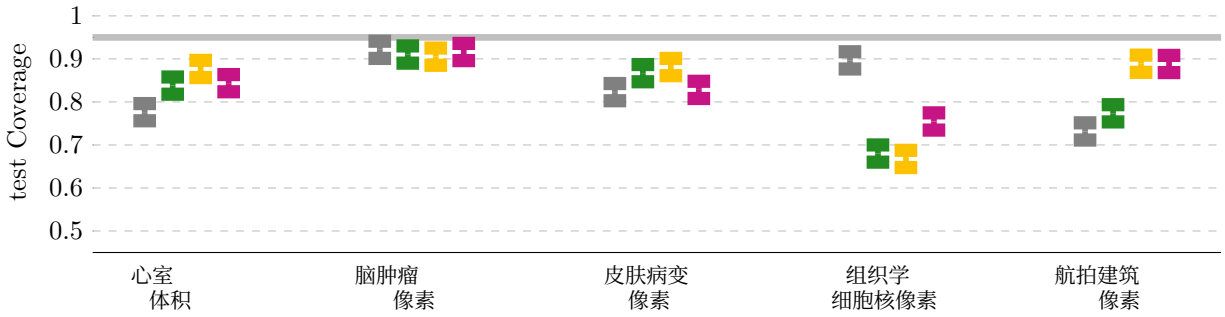


图 A12：失覆盖率 $\alpha = 0.05$：在五个真实世界数据集上，四种常见回归不确定性估计方法（保形预测、高斯、高斯集成、分位数回归）的测试集覆盖率结果。其他指标请参见表 A26–表 A30。

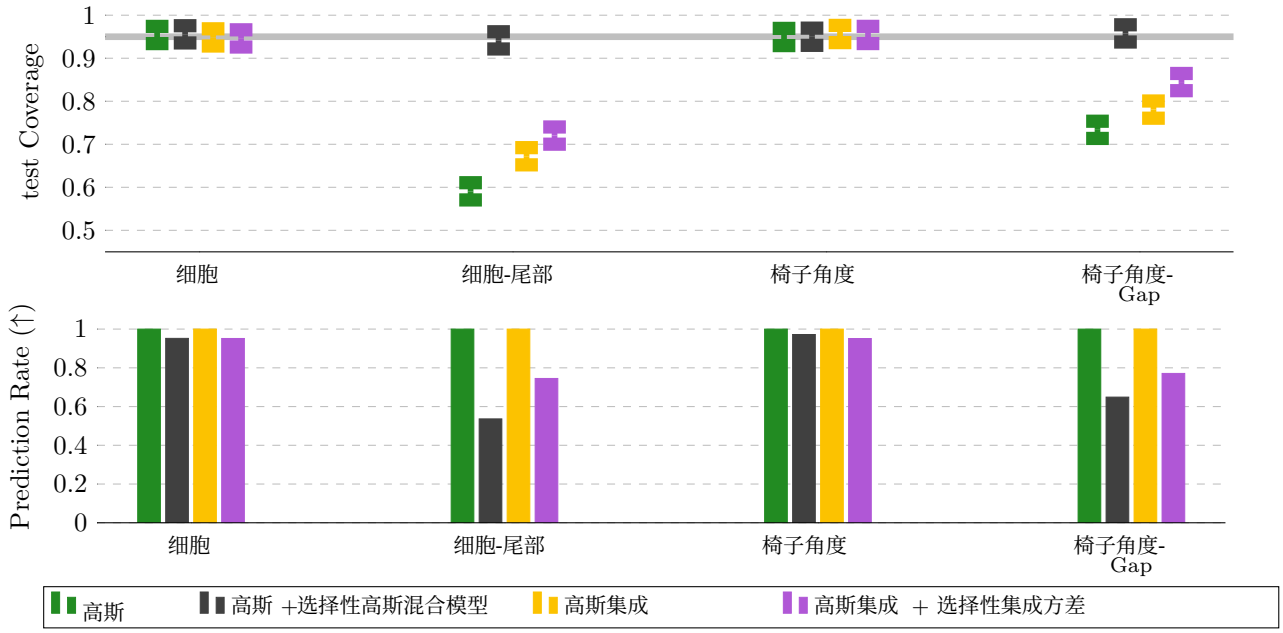


图 A13：失覆盖率 $\alpha = 0.05$：四个合成 数据集上两种选择性预测方法的结果。其他指标请参见表 A22 - 表 A25。

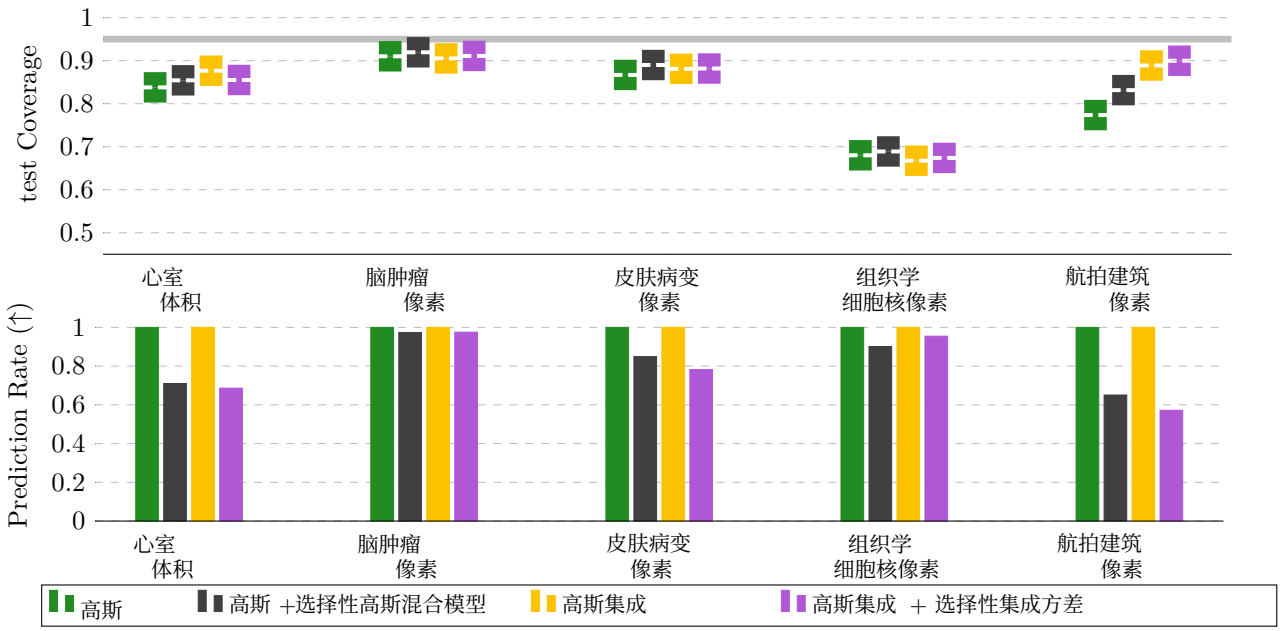


图 A14：失覆盖率 $\alpha = 0.05$：五个真实世界 数据集中的两个选择性预测方法的结果。其他指标见表 A26 - 表 A30。



图 A15：五种常见回归不确定性估计方法（保形预测、集成、高斯、高斯集成、分位数回归）在 AerialBuildingPixels 数据集及其两个测试集不同的版本上的结果。对于这三个数据集，训练/验证集均包含来自奥斯汀和芝加哥的图像。对于 AerialBuildingPixels，测试集包含来自奥地利西蒂罗尔的图像。对于 AerialBuildingPixels-Kitsap，测试集则包含来自美国华盛顿州基萨普县的图像。对于 AerialBuildingPixels-Vienna，测试集包含来自奥地利维也纳的图像。

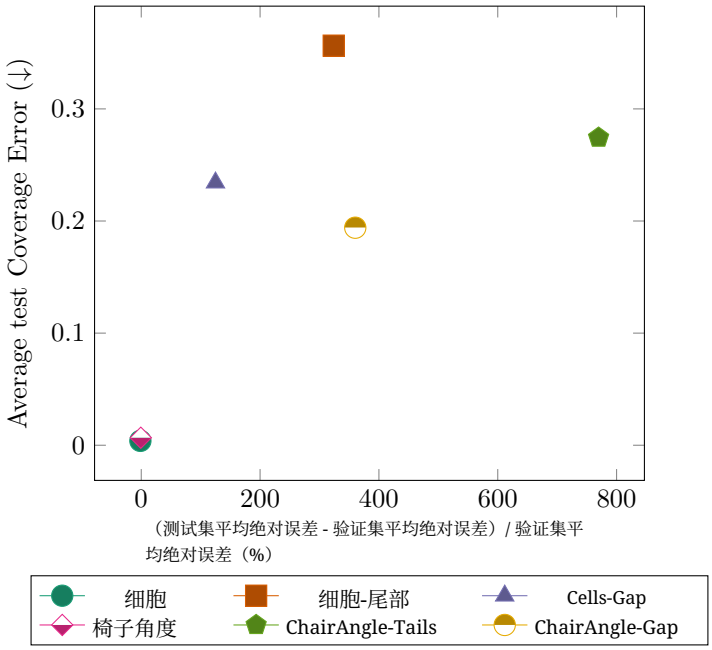


图 A16：使用验证集和测试集上的回归准确度差异（平均绝对误差）作为各数据集分布偏移的定量度量（灵感来自 Wenzel 等人（2022年），并扩展到我们的回归场景），并将其与测试集覆盖率性能进行比较。平均测试覆盖误差根据五种常见的回归不确定性估计方法计算。验证集/测试集平均绝对误差对应保形预测方法（标准直接回归模型）。六个合成数据集的结果。总体而言，分布偏移度量越大，测试集覆盖率性能越差。

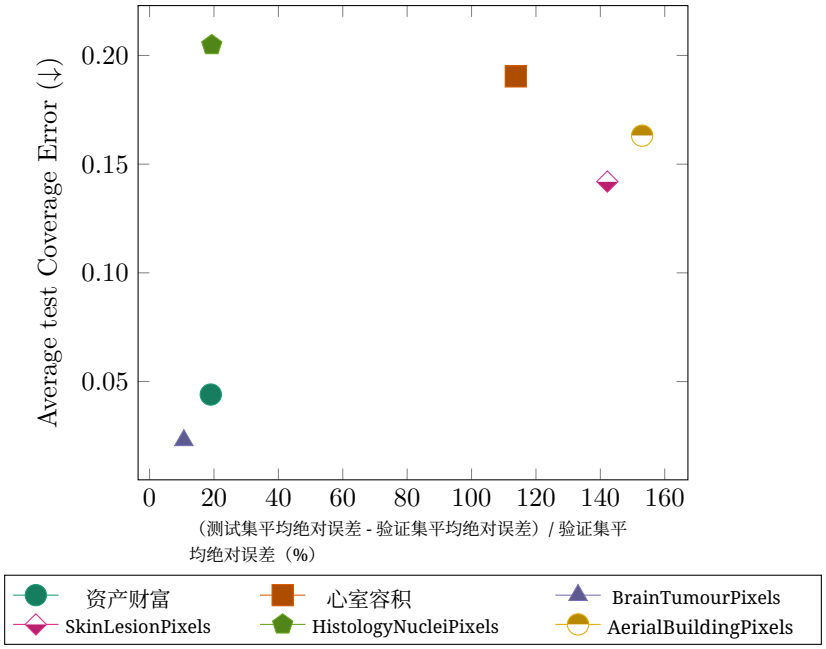


图 A17：与图 A16 完全相同的比较，但针对六个真实世界数据集。在这种情况下，较大的分布偏移度量通常也对应较差的测试覆盖率性能。HistologyNucleiPixels 数据集在一定程度上是个例外。

表A1：Cells 数据集上的完整结果。

方法	验证集平均绝对误差（↓）	验证集区间长度（↓）	测试集覆盖率（≥ 0.90）	测试集预测率 (↑)
保形预测	4.03121 ± 2.78375	18.8216 ± 11.8302	0.9117 ± 0.00275173	1.0
集成	2.22525 ± 0.471309	12.2881 ± 2.78795	0.8994 ± 0.00689986	1.0
高斯	3.61704 ± 1.13624	14.5492 ± 4.44927	0.90286 ± 0.00594629	1.0
高斯集成	2.78757 ± 1.16951	17.1285 ± 4.33367	0.90062 ± 0.0027571	1.0
分位数回归	3.70405 ± 0.81647	13.8023 ± 2.19085	0.90486 ± 0.006771	1.0
高斯 +选择性高斯混合模型	3.61704 ± 1.13624	14.5492 ± 4.44927	0.905241 ± 0.00448315	0.95216 ± 0.00290558
高斯 +选择性 k 近邻	3.61704 ± 1.13624	14.5492 ± 4.44927	0.900069 ± 0.00677637	0.94656 ± 0.00470472
高斯 +选择性方差	3.61704 ± 1.13624	14.5492 ± 4.44927	0.904839 ± 0.00795449	0.95712 ± 0.00200938
高斯集成 + 选择性高斯混合模型	2.78757 ± 1.16951	17.1285 ± 4.33367	0.899999 ± 0.00311328	0.95066 ± 0.0023105
高斯集成 + 选择性集成方差	2.78757 ± 1.16951	17.1285 ± 4.33367	0.896465 ± 0.0015705	0.95156 ± 0.00215277

表 A2：Cells-Tails 数据集上的完整结果。

方法	验证集平均绝对误差 (↓)	验证集区间长度 (↓)	测试集覆盖率 (≥ 0.90)	测试集预测率 (↑)
保形预测	3.56733 ± 1.0818	14.2128 ± 4.23582	0.55356 ± 0.0299096	1.0
集成	1.83461 ± 0.178937	9.8697 ± 1.6672	0.50078 ± 0.007608	1.0
高斯	4.05446 ± 1.33153	15.4321 ± 4.98796	0.54402 ± 0.0347525	1.0
高斯集成	2.40691 ± 0.580524	10.9696 ± 1.70051	0.5874 ± 0.0479263	1.0
分位数回归	3.32571 ± 1.24578	13.0848 ± 3.54239	0.52222 ± 0.0284102	1.0
高斯 +选择性高斯混合模型	4.05446 ± 1.33153	15.4321 ± 4.98796	0.889825 ± 0.0193021	0.53654 ± 0.0101012
高斯 +选择性 k 近邻	4.05446 ± 1.33153	15.4321 ± 4.98796	0.859179 ± 0.0255173	0.56692 ± 0.0127107
高斯 +选择性方差	4.05446 ± 1.33153	15.4321 ± 4.98796	0.687475 ± 0.0659807	0.6914 ± 0.0651196
高斯集成 + 选择性高斯混合模型	2.40691 ± 0.580524	10.9696 ± 1.70051	0.877068 ± 0.0222502	0.53604 ± 0.0125903
高斯集成 + 选择性集成方差	2.40691 ± 0.580524	10.9696 ± 1.70051	0.660212 ± 0.0325832	0.74512 ± 0.0438128

表 A3：Cells-Gap 数据集上的完整结果。

方法	验证集平均绝对误差 (↓)	验证集区间长度 (↓)	测试集覆盖率 (≥ 0.90)	测试集预测率 (↑)
保形预测	3.67702 ± 1.33587	17.1152 ± 6.49211	0.64002 ± 0.0620287	1.0
集成	2.7261 ± 0.705274	13.4061 ± 2.40513	0.6771 ± 0.0544788	1.0
高斯	3.53089 ± 1.0619	15.6396 ± 6.23458	0.66028 ± 0.114703	1.0
高斯集成	3.46118 ± 0.95429	18.707 ± 4.13287	0.7066 ± 0.0317638	1.0
分位数回归	4.75328 ± 1.73499	18.108 ± 4.21716	0.64488 ± 0.12875	1.0
高斯 + 选择性高斯混合模型	3.53089 ± 1.0619	15.6396 ± 6.23458	0.890569 ± 0.00953089	0.49372 ± 0.0025926
高斯 +选择性 k 近邻	3.53089 ± 1.0619	15.6396 ± 6.23458	0.874032 ± 0.0432364	0.53646 ± 0.0139864
高斯 +选择性方差	3.53089 ± 1.0619	15.6396 ± 6.23458	0.652766 ± 0.117192	0.9748 ± 0.000940213
高斯集成 + 选择性高斯混合模型	3.46118 ± 0.95429	18.707 ± 4.13287	0.896848 ± 0.0127879	0.49278 ± 0.00192914
高斯集成 + 选择性集成方差	3.46118 ± 0.95429	18.707 ± 4.13287	0.68326 ± 0.0299799	0.9147 ± 0.0182427

表 A4：ChairAngle 数据集上的完整结果。

方法	验证集平均绝对误差 (↓)	验证集区间长度 (↓)	测试集覆盖率 (≥ 0.90)	测试集预测率 (↑)
保形预测	0.289127 ± 0.081643	1.08752 ± 0.202247	0.905265 ± 0.00339915	1.0
集成	0.189858 ± 0.0548452	0.788401 ± 0.137465	0.909915 ± 0.00399101	1.0
高斯	0.376692 ± 0.171928	1.37757 ± 0.382191	0.901577 ± 0.00308472	1.0
高斯集成	0.361834 ± 0.165348	1.2044 ± 0.442422	0.910414 ± 0.00316257	1.0
分位数回归	0.851253 ± 0.544717	3.19741 ± 1.9751	0.906209 ± 0.00204038	1.0
高斯 +选择性高斯混合模型	0.376692 ± 0.171928	1.37757 ± 0.382191	0.902482 ± 0.00436054	0.972739 ± 0.00191733
高斯 +选择性 k 近邻	0.376692 ± 0.171928	1.37757 ± 0.382191	0.90222 ± 0.00459694	0.975465 ± 0.00201785
高斯 +选择性方差	0.376692 ± 0.171928	1.37757 ± 0.382191	0.904805 ± 0.00765098	0.961782 ± 0.0193133
高斯集成 + 选择性高斯混合模型	0.361834 ± 0.165348	1.2044 ± 0.442422	0.9093 ± 0.00297263	0.969033 ± 0.000940447
高斯集成 + 选择性集成方差	0.361834 ± 0.165348	1.2044 ± 0.442422	0.905905 ± 0.00336481	0.951359 ± 0.00453556

表 A5：ChairAngle-Tails 数据集上的完整结果。

方法	验证集平均绝对误差 (↓)	验证集区间长度 (↓)	测试集覆盖率 (≥ 0.90)	测试集预测率 (↑)
保形预测	0.358162 ± 0.168257	1.31102 ± 0.531401	0.615804 ± 0.00527695	1.0
集成	0.21931 ± 0.0596705	1.04074 ± 0.245846	0.611706 ± 0.00345611	1.0
高斯	0.241214 ± 0.091736	1.09417 ± 0.382907	0.622592 ± 0.00714065	1.0
高斯集成	0.13365 ± 0.0189933	0.769172 ± 0.0814039	0.617016 ± 0.0063698	1.0
分位数回归	0.820934 ± 0.653268	2.9815 ± 2.15473	0.660953 ± 0.0379624	1.0
高斯 +选择性高斯混合模型	0.241214 ± 0.091736	1.09417 ± 0.382907	0.901946 ± 0.00382993	0.655448 ± 0.001058
高斯 + 选择性 k 近邻	0.241214 ± 0.091736	1.09417 ± 0.382907	0.860311 ± 0.00617031	0.703038 ± 0.00691227
高斯 + 选择性方差	0.241214 ± 0.091736	1.09417 ± 0.382907	0.647559 ± 0.0360179	0.924383 ± 0.0751205
高斯集成 + 选择性高斯混合模型	0.13365 ± 0.0189933	0.769172 ± 0.0814039	0.904807 ± 0.00394199	0.651314 ± 0.0013893
高斯集成 + 选择性集成方差	0.13365 ± 0.0189933	0.769172 ± 0.0814039	0.751845 ± 0.0131765	0.772401 ± 0.00878839

表 A6：ChairAngle-Gap 数据集上的完整结果。

方法	验证集平均绝对误差 (↓)	验证集区间长度 (↓)	测试集覆盖率 (≥ 0.90)	测试集预测率 (↑)
保形预测	0.35034 ± 0.161819	1.35 ± 0.547089	0.659546 ± 0.00916108	1.0
集成	0.22898 ± 0.0853825	1.1222 ± 0.174147	0.749951 ± 0.0155327	1.0
高斯	0.454516 ± 0.280174	2.09212 ± 0.933756	0.69363 ± 0.0345876	1.0
高斯集成	0.226352 ± 0.0677413	1.30588 ± 0.136235	0.731065 ± 0.0126216	1.0
分位数回归	0.639151 ± 0.296536	3.29137 ± 1.53269	0.695697 ± 0.0413613	1.0
高斯 +选择性高斯混合模型	0.454516 ± 0.280174	2.09212 ± 0.933756	0.91215 ± 0.00604745	0.649372 ± 0.00334919
高斯 +选择性 k 近邻	0.454516 ± 0.280174	2.09212 ± 0.933756	0.911574 ± 0.00376608	0.673764 ± 0.0113432
高斯 +选择性方差	0.454516 ± 0.280174	2.09212 ± 0.933756	0.690436 ± 0.0366025	0.981292 ± 0.0152406
高斯集成 + 选择性高斯混合模型	0.226352 ± 0.0677413	1.30588 ± 0.136235	0.905039 ± 0.00229837	0.648624 ± 0.00094348
高斯集成 + 选择性集成方差	0.226352 ± 0.0677413	1.30588 ± 0.136235	0.794531 ± 0.014876	0.7711 ± 0.0191172

表 A7：AssetWealth 数据集上的完整结果。

方法	验证集平均绝对误差 (↓)	验证集区间长度（↓）	测试集覆盖率（≥ 0.90）	测试集预测率（↑）
保形预测	0.346532 ± 0.00578306	1.5838 ± 0.0318086	0.87136 ± 0.0090944	1.0
集成	0.320002 ± 0.00264202	1.45568 ± 0.0129312	0.87348 ± 0.00598963	1.0
高斯	0.367501 ± 0.0416437	1.597 ± 0.207599	0.844966 ± 0.0456831	1.0
高斯集成	0.3295 ± 0.00783906	1.42071 ± 0.0485437	0.866162 ± 0.00711672	1.0
分位数回归	0.404279 ± 0.0683226	1.58957 ± 0.161689	0.823921 ± 0.0313343	1.0
高斯 +选择性高斯混合模型	0.367501 ± 0.0416437	1.597 ± 0.207599	0.850824 ± 0.047533	0.93838 ± 0.0170443
高斯 +选择性 k 近邻	0.367501 ± 0.0416437	1.597 ± 0.207599	0.852107 ± 0.0474734	0.933586 ± 0.0203541
高斯 +选择性方差	0.367501 ± 0.0416437	1.597 ± 0.207599	0.846981 ± 0.0430247	0.926874 ± 0.0159098
高斯集成 + 选择性高斯混合模型	0.3295 ± 0.00783906	1.42071 ± 0.0485437	0.868444 ± 0.00727312	0.937371 ± 0.0138213
高斯集成 + 选择性集成方差	0.3295 ± 0.00783906	1.42071 ± 0.0485437	0.85047 ± 0.00397082	0.868282 ± 0.0299149

表 A8：VentricularVolume 数据集上的完整结果。

方法	验证集平均绝对误差 (↓)	验证集区间长度 (↓)	测试集覆盖率 (≥ 0.90)	测试集预测率（↑）
保形预测	11.2471 ± 0.201399	47.4505 ± 1.556	0.603135 ± 0.0125842	1.0
集成	10.2476 ± 0.113707	41.8445 ± 0.750822	0.707367 ± 0.00877743	1.0
高斯	12.7238 ± 1.52197	51.566 ± 3.52739	0.730878 ± 0.0619045	1.0
高斯集成	10.1141 ± 0.180661	39.9817 ± 1.91308	0.795768 ± 0.0231243	1.0
分位数回归	12.4944 ± 0.676265	49.0448 ± 3.33314	0.710972 ± 0.045171	1.0
高斯 +选择性高斯混合模型	12.7238 ± 1.52197	51.566 ± 3.52739	0.752046 ± 0.0529087	0.707994 ± 0.0208741
高斯 +选择性 k 近邻	12.7238 ± 1.52197	51.566 ± 3.52739	0.735105 ± 0.0612413	0.911599 ± 0.0447475
高斯 +选择性方差	12.7238 ± 1.52197	51.566 ± 3.52739	0.747868 ± 0.0569984	0.691693 ± 0.0341016
高斯集成 + 选择性高斯混合模型	10.1141 ± 0.180661	39.9817 ± 1.91308	0.798094 ± 0.0166674	0.646865 ± 0.0366201
高斯集成 + 选择性集成方差	10.1141 ± 0.180661	39.9817 ± 1.91308	0.763412 ± 0.0286772	0.686207 ± 0.0431835

表 A9：BrainTumourPixels 数据集上的完整结果。

方法	验证集平均绝对误差 (↓)	验证集区间长度 (↓)	测试集覆盖率 (≥ 0.90)	测试集预测率 (↑)
保形预测	21.3163 ± 0.45997	93.7169 ± 2.99061	0.885925 ± 0.00395743	1.0
集成	21.1133 ± 0.210209	90.3825 ± 0.825858	0.878183 ± 0.00318872	1.0
高斯	21.0625 ± 0.358012	93.6284 ± 2.29916	0.873544 ± 0.00871896	1.0
高斯集成	20.5336 ± 0.211421	87.7414 ± 0.818979	0.868426 ± 0.00328047	1.0
分位数回归	22.0348 ± 0.697606	94.3249 ± 3.07265	0.879079 ± 0.00380396	1.0
高斯 +选择性高斯混合模型	21.0625 ± 0.358012	93.6284 ± 2.29916	0.883515 ± 0.0138279	0.973576 ± 0.0187252
高斯 +选择性 k 近邻	21.0625 ± 0.358012	93.6284 ± 2.29916	0.891264 ± 0.00734602	0.947185 ± 0.0178434
高斯 +选择性方差	21.0625 ± 0.358012	93.6284 ± 2.29916	0.878666 ± 0.00735197	0.978791 ± 0.00694672
高斯集成 + 选择性高斯混合模型	20.5336 ± 0.211421	87.7414 ± 0.818979	0.873824 ± 0.00466519	0.985349 ± 0.00461984
高斯集成 + 选择性集成方差	20.5336 ± 0.211421	87.7414 ± 0.818979	0.877211 ± 0.00302057	0.975368 ± 0.00323397

表A10：SkinLesionPixels 数据集上的完整结果。

方法	验证集平均绝对误差（↓）	验证集区间长度（↓）	测试集覆盖率（≥ 0.90）	测试集预测率 (↑)
保形预测	107.514 ± 1.87464	492.922 ± 14.1865	0.708012 ± 0.00917777	1.0
集成	99.3156 ± 0.997867	353.842 ± 4.73577	0.742098 ± 0.00852236	1.0
高斯	105.417 ± 1.15178	535.139 ± 215.446	0.797255 ± 0.0270734	1.0
高斯集成	100.639 ± 0.464183	472.14 ± 85.2654	0.822931 ± 0.0134328	1.0
分位数回归	113.076 ± 3.20875	405.904 ± 4.75158	0.719788 ± 0.0213826	1.0
高斯 +选择性高斯混合模型	105.417 ± 1.15178	535.139 ± 215.446	0.821515 ± 0.0137705	0.849579 ± 0.0206181
高斯 +选择性 k 近邻	105.417 ± 1.15178	535.139 ± 215.446	0.813027 ± 0.0306586	0.927047 ± 0.00830718
高斯 +选择性方差	105.417 ± 1.15178	535.139 ± 215.446	0.799821 ± 0.0150855	0.77946 ± 0.059169
高斯集成 + 选择性高斯混合模型	100.639 ± 0.464183	472.14 ± 85.2654	0.826921 ± 0.00719649	0.839132 ± 0.0108815
高斯集成 + 选择性集成方差	100.639 ± 0.464183	472.14 ± 85.2654	0.819129 ± 0.00910128	0.782116 ± 0.0109776

表 A11：HistologyNucleiPixels 数据集上的完整结果。

方法	验证集平均绝对误差 (↓)	验证集区间长度 (↓)	测试集覆盖率 (≥ 0.90)	测试集预测率 (↑)
保形预测	217.887 ± 3.71766	993.665 ± 18.3282	0.841288 ± 0.0128314	1.0
集成	197.078 ± 1.99489	853.361 ± 14.1904	0.812704 ± 0.00664074	1.0
高斯	211.795 ± 10.0239	1211.83 ± 396.946	0.588796 ± 0.0912794	1.0
高斯集成	196.785 ± 2.14454	1108.53 ± 145.034	0.54936 ± 0.05995	1.0
分位数回归	227.895 ± 9.37415	914.909 ± 45.995	0.684076 ± 0.0428704	1.0
高斯 + 选择性高斯混合模型	211.795 ± 10.0239	1211.83 ± 396.946	0.59554 ± 0.0956742	0.90569 ± 0.0453555
高斯 +选择性 k 近邻	211.795 ± 10.0239	1211.83 ± 396.946	0.608929 ± 0.0805954	0.846228 ± 0.0482832
高斯 +选择性方差	211.795 ± 10.0239	1211.83 ± 396.946	0.588598 ± 0.0913593	0.998765 ± 0.00134954
高斯集成 + 选择性高斯混合模型	196.785 ± 2.14454	1108.53 ± 145.034	0.553725 ± 0.0642274	0.882488 ± 0.0156395
高斯集成 + 选择性集成方差	196.785 ± 2.14454	1108.53 ± 145.034	0.555149 ± 0.0620052	0.954477 ± 0.0256246

表A12：AerialBuildingPixels 数据集上的完整结果。

方法	验证集平均绝对误差 (↓)	验证集区间长度 (↓)	测试集覆盖率 (≥ 0.90)	测试集预测率 (↑)
保形预测	235.417 ± 7.16096	1038.45 ± 51.5565	0.637584 ± 0.0691102	1.0
集成	199.206 ± 2.76543	798.312 ± 6.80986	0.772494 ± 0.0227288	1.0
高斯	217.877 ± 1.72493	929.562 ± 47.6606	0.698766 ± 0.0662263	1.0
高斯集成	208.487 ± 1.03581	885.349 ± 27.8584	0.812339 ± 0.0617386	1.0
分位数回归	242.284 ± 6.0122	909.191 ± 24.9528	0.763342 ± 0.0902358	1.0
高斯 + 选择性高斯混合模型	217.877 ± 1.72493	929.562 ± 47.6606	0.76535 ± 0.0388677	0.652082 ± 0.0990489
高斯 +选择性 k 近邻	217.877 ± 1.72493	929.562 ± 47.6606	0.651867 ± 0.0771154	0.840103 ± 0.0463989
高斯 +选择性方差	217.877 ± 1.72493	929.562 ± 47.6606	0.725714 ± 0.0779575	0.708226 ± 0.103803
高斯集成 + 选择性高斯混合模型	208.487 ± 1.03581	885.349 ± 27.8584	0.847101 ± 0.038266	0.686787 ± 0.0278702
高斯集成 + 选择性集成方差	208.487 ± 1.03581	885.349 ± 27.8584	0.838352 ± 0.0312236	0.571928 ± 0.0402727

表 A13：失覆盖率 $\alpha = 0.2$：Cells 数据集上的结果。

方法	验证集平均绝对误差 (↓)	验证集区间长度 (↓)	测试集覆盖率 (≥ 0.80)	测试集预测率（↑）
保形预测	4.03121 ± 2.78375	14.042 ± 10.2791	0.811 ± 0.0100485	1.0
高斯	3.61704 ± 1.13624	11.544 ± 3.59513	0.808 ± 0.00859046	1.0
高斯集成	2.78757 ± 1.16951	13.1298 ± 3.34997	0.80044 ± 0.00316582	1.0
分位数回归	4.40758 ± 1.47411	13.3314 ± 4.39511	0.80152 ± 0.00667005	1.0
高斯 + 选择性高斯混合模型	3.61704 ± 1.13624	11.544 ± 3.59513	0.811076 ± 0.00507358	0.95216 ± 0.00280043
高斯集成 + 选择性集成方差	2.78757 ± 1.16951	13.1298 ± 3.34997	0.793124 ± 0.00576659	0.95156 ± 0.00215277

表 A14：失覆盖率 $\alpha = 0.2$：Cells-Tails 数据集上的结果。

方法	验证集平均绝对误差 (↓)	验证集区间长度 (↓)	测试集覆盖率 (≥ 0.80)	测试集预测率（↑）
保形预测	3.56733 ± 1.0818	11.3841 ± 3.54242	0.48332 ± 0.0337429	1.0
高斯	4.05446 ± 1.33153	12.7086 ± 4.39552	0.47106 ± 0.0319074	1.0
高斯集成	2.40691 ± 0.580524	8.62187 ± 1.5177	0.4768 ± 0.0461347	1.0
分位数回归	4.20398 ± 1.61554	13.1241 ± 4.61204	0.49358 ± 0.0341114	1.0
高斯 +选择性高斯混合模型	4.05446 ± 1.33153	12.7086 ± 4.39552	0.788228 ± 0.0239896	0.5364 ± 0.0100584
高斯集成 + 选择性集成方差	2.40691 ± 0.580524	8.62187 ± 1.5177	0.567734 ± 0.0270648	0.74512 ± 0.0438128

表A15：失覆盖率 $\alpha = 0.2$：ChairAngle 数据集上的结果。

方法	验证集平均绝对误差 (↓)	验证集区间长度 (↓)	测试集覆盖率 (≥ 0.80)	测试集预测率 (↑)
保形预测	0.289127 ± 0.081643	0.872808 ± 0.205426	0.80212 ± 0.00337798	1.0
高斯	0.376692 ± 0.171928	1.14193 ± 0.40832	0.806646 ± 0.00565869	1.0
高斯集成	0.361834 ± 0.165348	1.01037 ± 0.399644	0.813452 ± 0.00329742	1.0
分位数回归	0.648062 ± 0.181932	1.87322 ± 0.428951	0.807127 ± 0.00380433	1.0
高斯 +选择性高斯混合模型	0.376692 ± 0.171928	1.14193 ± 0.40832	0.807801 ± 0.0066405	0.972829 ± 0.0019256
高斯集成 + 选择性集成方差	0.361834 ± 0.165348	1.01037 ± 0.399644	0.804719 ± 0.00318856	0.951359 ± 0.00453556

表 A16：失覆盖率 $\alpha = 0.2$：ChairAngle-Gap 数据集上的结果。

方法	验证集平均绝对误差 (↓)	验证集区间长度 (↓)	测试集覆盖率 (≥ 0.80)	测试集预测率 (↑)
保形预测	0.35034 ± 0.161819	1.09877 ± 0.497591	0.588472 ± 0.0119766	1.0
高斯	0.454516 ± 0.280174	1.69802 ± 0.840083	0.613951 ± 0.0358797	1.0
高斯集成	0.226352 ± 0.0677413	0.972076 ± 0.118477	0.6431 ± 0.0136145	1.0
分位数回归	0.692597 ± 0.269476	2.29958 ± 0.738934	0.600606 ± 0.0136385	1.0
高斯 +选择性高斯混合模型	0.454516 ± 0.280174	1.69802 ± 0.840083	0.813442 ± 0.00720691	0.650227 ± 0.00138861
高斯集成 + 选择性集成方差	0.226352 ± 0.0677413	0.972076 ± 0.118477	0.701301 ± 0.0146886	0.7711 ± 0.0191172

表 A17：失覆盖率 $\alpha = 0.2$：心室容积数据集上的结果 et.

方法	验证集平均绝对误差 (↓)	验证集区间长度 (↓)	测试集覆盖率 (≥ 0.80)	测试集预测率（↑）
保形预测	11.2471 ± 0.201399	33.0996 ± 0.991594	0.457053 ± 0.00583106	1.0
高斯	12.7238 ± 1.52197	37.3172 ± 4.91536	0.589342 ± 0.0793156	1.0
高斯集成	10.1141 ± 0.180661	29.9914 ± 1.12007	0.678527 ± 0.0277009	1.0
分位数回归	12.5465 ± 0.941364	35.8073 ± 1.5139	0.537774 ± 0.0444887	1.0
高斯 +选择性高斯混合模型	12.7238 ± 1.52197	37.3172 ± 4.91536	0.609177 ± 0.0682058	0.709875 ± 0.0232873
高斯集成 + 选择性集成方差	10.1141 ± 0.180661	29.9914 ± 1.12007	0.640078 ± 0.0277429	0.686207 ± 0.0431835

表 A18：失覆盖 率 $\alpha = 0.2$：BrainTum ourPixels 数据集上的结果。

方法	验证集平均绝对误差（↓）	验证集区间长度 (↓)	测试集覆盖率 (≥ 0.80)	测试集预测率 (↑)
保形预测	21.3163 ± 0.45997	66.146 ± 2.75637	0.824632 ± 0.0128289	1.0
高斯	21.0625 ± 0.358012	64.2423 ± 1.9975	0.798113 ± 0.0133438	1.0
高斯集成	20.5336 ± 0.211421	62.5706 ± 0.713494	0.796961 ± 0.00658989	1.0
分位数回归	21.9545 ± 0.149787	66.2654 ± 1.33091	0.807646 ± 0.0165337	1.0
高斯 +选择性高斯混合模型	21.0625 ± 0.358012	64.2423 ± 1.9975	0.809477 ± 0.0139729	0.973576 ± 0.0187252
高斯集成 + 选择性集成方差	20.5336 ± 0.211421	62.5706 ± 0.713494	0.806998 ± 0.00641546	0.975368 ± 0.00323397

表 A19：失覆盖率 $\alpha = 0.2$：SkinLesionPixels 数据集上的结果。

方法	验证集平均绝对误差 (↓)	验证集区间长度 (↓)	测试集覆盖率 (≥ 0.80)	测试集预测率 (↑)
保形预测	107.514 ± 1.87464	275.494 ± 5.60123	0.553254 ± 0.0122434	1.0
高斯	105.417 ± 1.15178	395.661 ± 174.897	0.689951 ± 0.0406148	1.0
高斯集成	100.639 ± 0.464183	351.961 ± 70.4308	0.715272 ± 0.0278757	1.0
分位数回归	106.382 ± 2.70593	251.346 ± 8.67954	0.590792 ± 0.0115523	1.0
高斯 +选择性高斯混合模型	105.417 ± 1.15178	395.661 ± 174.897	0.708767 ± 0.0274691	0.849668 ± 0.0189464
高斯集成 + 选择性集成方差	100.639 ± 0.464183	351.961 ± 70.4308	0.704273 ± 0.0198608	0.782116 ± 0.0109776

Table A20：失覆盖率 $\alpha = 0.2$：HistologyNucleiPixels 数据集上的结果 .

方法	验证集平均绝对误差 (↓)	验证集区间长度 (↓)	测试集覆盖率 (≥ 0.80)	测试集预测率 (↑)
保形预测	217.887 ± 3.71766	688.269 ± 13.6761	0.744155 ± 0.0107073	1.0
高斯	211.795 ± 10.0239	902.884 ± 295.983	0.454345 ± 0.0986975	1.0
高斯集成	196.785 ± 2.14454	840.287 ± 113.005	0.412528 ± 0.0544736	1.0
分位数回归	221.136 ± 9.7918	679.117 ± 30.6895	0.580503 ± 0.0463751	1.0
高斯 + 选择性高斯混合模型	211.795 ± 10.0239	902.884 ± 295.983	0.462248 ± 0.105935	0.901985 ± 0.0400294
高斯集成 + 选择性集成方差	196.785 ± 2.14454	840.287 ± 113.005	0.417977 ± 0.0572495	0.954477 ± 0.0256246

表 A21：失覆盖率 $\alpha = 0.2$：AerialBuildingPixels 数据集上的结果 et.

方法	验证集平均绝对误差（↓）	验证集区间长度（↓）	测试集覆盖率（≥ 0.80）	测试集预测率 (↑)
保形预测	235.417 ± 7.16096	685.829 ± 31.3669	0.52437 ± 0.0610316	1.0
高斯	217.877 ± 1.72493	685.899 ± 43.1749	0.60129 ± 0.0717427	1.0
高斯集成	208.487 ± 1.03581	658.243 ± 27.1206	0.683599 ± 0.0674139	1.0
分位数回归	229.243 ± 6.21038	662.104 ± 18.5495	0.635733 ± 0.0350343	1.0
高斯 + 选择性高斯混合模型	217.877 ± 1.72493	685.899 ± 43.1749	0.679246 ± 0.0462652	0.65018 ± 0.0999173
高斯集成 + 选择性集成方差	208.487 ± 1.03581	658.243 ± 27.1206	0.743714 ± 0.0354023	0.571928 ± 0.0402727

表 A22：失覆盖率 $\alpha = 0.05$：Cells 数据集上的结果。

方法	验证集平均绝对误差 (↓)	验证集区间长度 (↓)	测试集覆盖率 (≥ 0.95)	测试集预测率 (↑)
保形预测	4.03121 ± 2.78375	22.3973 ± 12.7194	0.95612 ± 0.00460235	1.0
高斯	3.61704 ± 1.13624	17.0459 ± 4.98381	0.95402 ± 0.00275928	1.0
高斯集成	2.78757 ± 1.16951	20.5445 ± 5.00176	0.9484 ± 0.00331843	1.0
分位数回归	3.5478 ± 1.25834	17.8244 ± 3.57608	0.94858 ± 0.00565346	1.0
高斯 + 选择性高斯混合模型	3.61704 ± 1.13624	17.0459 ± 4.98381	0.955558 ± 0.0032579	0.95212 ± 0.00276362
高斯集成 + 选择性集成方差	2.78757 ± 1.16951	20.5445 ± 5.00176	0.946085 ± 0.00337562	0.95156 ± 0.00215277

表 A23：失覆盖率 $\alpha = 0.05$：Cells-Tails 数据集上的结果。

方法	验证集平均绝对误差 (↓)	验证集区间长度 (↓)	测试集覆盖率 (≥ 0.95)	测试集预测率 (↑)
保形预测	3.56733 ± 1.0818	16.3831 ± 4.80611	0.59048 ± 0.0307226	1.0
高斯	4.05446 ± 1.33153	17.5443 ± 5.3978	0.59066 ± 0.0417946	1.0
高斯集成	2.40691 ± 0.580524	13.1205 ± 1.81604	0.67218 ± 0.0308246	1.0
分位数回归	3.34375 ± 0.642763	14.5539 ± 2.01592	0.5666 ± 0.0295055	1.0
高斯 +选择性高斯混合模型	4.05446 ± 1.33153	17.5443 ± 5.3978	0.941551 ± 0.0164539	0.53636 ± 0.0100057
高斯集成 + 选择性集成方差	2.40691 ± 0.580524	13.1205 ± 1.81604	0.720221 ± 0.0374625	0.74512 ± 0.0438128

表 A24：失覆盖率 $\alpha = 0.05$：ChairAngle 数据集上的结果。

方法	验证集平均绝对误差 (↓)	验证集区间长度 (↓)	测试集覆盖率 (≥ 0.95)	测试集预测率 (↑)
保形预测	0.289127 ± 0.081643	1.27437 ± 0.18759	0.953087 ± 0.000729425	1.0
高斯	0.376692 ± 0.171928	1.59112 ± 0.367248	0.949381 ± 0.00143361	1.0
高斯集成	0.361834 ± 0.165348	1.38384 ± 0.481345	0.95633 ± 0.0016841	1.0
分位数回归	0.888707 ± 0.472522	4.04185 ± 1.50356	0.951608 ± 0.00184683	1.0
高斯 + 选择性高斯混合模型	0.376692 ± 0.171928	1.59112 ± 0.367248	0.949963 ± 0.00265095	0.97208 ± 0.00166628
高斯集成 + 选择性集成方差	0.361834 ± 0.165348	1.38384 ± 0.481345	0.9541 ± 0.00167267	0.951359 ± 0.00453556

表 A25：失覆盖率 $\alpha = 0.05$：ChairAngle-Gap 数据集上的结果。

方法	验证集平均绝对误差 (↓)	验证集区间长度 (↓)	测试集覆盖率 (≥ 0.95)	测试集预测率（↑）
保形预测	0.35034 ± 0.161819	1.56264 ± 0.582954	0.697817 ± 0.00918537	1.0
高斯	0.454516 ± 0.280174	2.39892 ± 1.02472	0.733595 ± 0.0367327	1.0
高斯集成	0.226352 ± 0.0677413	1.59703 ± 0.159795	0.780615 ± 0.0141415	1.0
分位数回归	1.43807 ± 0.99796	5.57113 ± 1.71783	0.759929 ± 0.034465	1.0
高斯 +选择性高斯混合模型	0.454516 ± 0.280174	2.39892 ± 1.02472	0.95774 ± 0.00281998	0.64882 ± 0.00276083
高斯集成 + 选择性集成方差	0.226352 ± 0.0677413	1.59703 ± 0.159795	0.844587 ± 0.0165508	0.7711 ± 0.0191172

表 A26：失覆盖率 $\alpha = 0.05$：VentricularVolume 数据集上的结果 t.

方法	验证集平均绝对误差 (↓)	验证集区间长度 (↓)	测试集覆盖率 (≥ 0.95)	测试集预测率 (↑)
保形预测	11.2471 ± 0.201399	70.211 ± 1.37665	0.776176 ± 0.00853621	1.0
高斯	12.7238 ± 1.52197	67.1562 ± 6.31002	0.837931 ± 0.0396691	1.0
高斯集成	10.1141 ± 0.180661	51.9065 ± 3.94789	0.876646 ± 0.0140944	1.0
分位数回归	12.0793 ± 0.130032	61.623 ± 3.04835	0.843574 ± 0.0243659	1.0
高斯 +选择性高斯混合模型	12.7238 ± 1.52197	67.1562 ± 6.31002	0.854302 ± 0.0279176	0.710188 ± 0.0225912
高斯集成 + 选择性集成方差	10.1141 ± 0.180661	51.9065 ± 3.94789	0.855053 ± 0.0125036	0.686207 ± 0.0431835

表 A27：失覆盖率 $\alpha = 0.05$：BrainTumourPixels 数据集上的结果 et.

方法	验证集平均绝对误差 (↓)	验证集区间长度 (↓)	测试集覆盖率 (≥ 0.95)	测试集预测率 (↑)
保形预测	21.3163 ± 0.45997	123.719 ± 2.07495	0.921113 ± 0.00123648	1.0
高斯	21.0625 ± 0.358012	123.587 ± 3.3168	0.910141 ± 0.00537523	1.0
高斯集成	20.5336 ± 0.211421	113.592 ± 1.37398	0.905182 ± 0.00359141	1.0
分位数回归	24.6897 ± 2.49388	126.788 ± 7.00165	0.915995 ± 0.00956349	1.0
高斯 +选择性高斯混合模型	21.0625 ± 0.358012	123.587 ± 3.3168	0.919455 ± 0.0105219	0.973672 ± 0.0186442
高斯集成 + 选择性集成方差	20.5336 ± 0.211421	113.592 ± 1.37398	0.910995 ± 0.00267557	0.975368 ± 0.00323397

表 A28：失覆盖率 $\alpha = 0.05$：SkinLesionPixels 数据集上的结果 .

方法	验证集平均绝对误差 (↓)	验证集区间长度 (↓)	测试集覆盖率 (≥ 0.95)	测试集预测率 (↑)
保形预测	107.514 ± 1.87464	845.109 ± 31.2572	0.822753 ± 0.0104666	1.0
高斯	105.417 ± 1.15178	691.768 ± 232.071	0.866844 ± 0.0166304	1.0
高斯集成	100.639 ± 0.464183	582.719 ± 95.3488	0.881009 ± 0.00686587	1.0
分位数回归	112.07 ± 5.13986	627.938 ± 68.3371	0.827977 ± 0.017595	1.0
高斯 +选择性高斯混合模型	105.417 ± 1.15178	691.768 ± 232.071	0.889891 ± 0.00940058	0.849225 ± 0.0194809
高斯集成 + 选择性集成方差	100.639 ± 0.464183	582.719 ± 95.3488	0.881749 ± 0.00534917	0.782116 ± 0.0109776

Ta表 A29：失覆盖率 $\alpha = 0.05$：HistologyNucleiPixels 数据集上的结果 .

方法	验证集平均绝对误差 (↓)	验证集区间长度 (↓)	测试集覆盖率 (≥ 0.95)	测试集预测率 (↑)
保形预测	217.887 ± 3.71766	1282.78 ± 25.9915	0.896603 ± 0.0113207	1.0
高斯	211.795 ± 10.0239	1491.62 ± 472.378	0.680018 ± 0.0860417	1.0
高斯集成	196.785 ± 2.14454	1358.12 ± 174.189	0.66749 ± 0.0522544	1.0
分位数回归	251.562 ± 12.244	1191.96 ± 65.3122	0.754566 ± 0.0379023	1.0
高斯 + 选择性高斯混合模型	211.795 ± 10.0239	1491.62 ± 472.378	0.68897 ± 0.0895173	0.901191 ± 0.0405698
高斯集成 + 选择性集成方差	196.785 ± 2.14454	1358.12 ± 174.189	0.673945 ± 0.0533738	0.954477 ± 0.0256246

表 A30：失覆盖率 $\alpha = 0.05$：AerialBuildingPixels 数据集上的结果。

方法	验证集平均绝对误差 (↓)	验证集区间长度 (↓)	测试集覆盖率 (≥ 0.95)	测试集预测率 (↑)
保形预测	235.417 ± 7.16096	1474.93 ± 53.6301	0.731414 ± 0.070291	1.0
高斯	217.877 ± 1.72493	1181.55 ± 41.5859	0.773522 ± 0.0548329	1.0
高斯集成	208.487 ± 1.03581	1108.87 ± 21.3551	0.88874 ± 0.0490213	1.0
分位数回归	294.181 ± 19.9268	1281.37 ± 51.1001	0.888175 ± 0.028968	1.0
高斯 + 选择性高斯混合模型	217.877 ± 1.72493	1181.55 ± 41.5859	0.831501 ± 0.0374173	0.650797 ± 0.0987431
高斯集成 + 选择性集成方差	208.487 ± 1.03581	1108.87 ± 21.3551	0.899838 ± 0.0263409	0.571928 ± 0.0402727

表 A31：Cells 数据集上的方法变体结果。

方法	验证集平均绝对误差 (↓)	验证集区间长度 (↓)	测试集覆盖率 (≥ 0.90)	测试集预测率（↑）
高斯 +选择性高斯混合模型	3.61704 ± 1.13624	14.5492 ± 4.44927	0.905241 ± 0.00448315	0.95216 ± 0.00290558
高斯 + 选择性高斯混合模型， $k = 2$	3.61704 ± 1.13624	14.5492 ± 4.44927	0.9044 ± 0.00481955	0.952 ± 0.00270259
高斯 + 选择性高斯混合模型， $k = 8$	3.61704 ± 1.13624	14.5492 ± 4.44927	0.904928 ± 0.00469093	0.9528 ± 0.00331662
高斯 + 选择性高斯混合模型，谱范数	4.80289 ± 1.89759	18.5596 ± 9.13391	0.908 ± 0.00703097	0.95118 ± 0.00220672
高斯 + 选择性 k 近邻	3.61704 ± 1.13624	14.5492 ± 4.44927	0.900069 ± 0.00677637	0.94656 ± 0.00470472
高斯 + 选择性 k 近邻， $k = 5$	3.61704 ± 1.13624	14.5492 ± 4.44927	0.900146 ± 0.00671325	0.94736 ± 0.00395049
高斯 + 选择性 k 近邻， $k = 20$	3.61704 ± 1.13624	14.5492 ± 4.44927	0.900048 ± 0.00691667	0.94624 ± 0.00248564
高斯 + 选择性 k 近邻，L2范数	3.61704 ± 1.13624	14.5492 ± 4.44927	0.906295 ± 0.00579854	0.95166 ± 0.00422734

表 A32：Cells-Tails 数据集上的方法变体结果。

方法	验证集平均绝对误差（↓）	验证集区间长度 (↓)	测试集覆盖率 (≥ 0.90)	测试集预测率 (↑)
高斯 +选择性高斯混合模型	4.05446 ± 1.33153	15.4321 ± 4.98796	0.889825 ± 0.0193021	0.53654 ± 0.0101012
高斯 + 选择性高斯混合模型， $k = 2$	4.05446 ± 1.33153	15.4321 ± 4.98796	0.891933 ± 0.0175265	0.525 ± 0.00829409
高斯 + 选择性高斯混合模型， $k = 8$	4.05446 ± 1.33153	15.4321 ± 4.98796	0.893648 ± 0.0182098	0.53516 ± 0.00975635
高斯 + 选择性高斯混合模型，谱范数	4.91114 ± 2.46961	17.3945 ± 7.3177	0.881351 ± 0.021587	0.5331 ± 0.0157885
高斯 + 选择性 k 近邻	4.05446 ± 1.33153	15.4321 ± 4.98796	0.859179 ± 0.0255173	0.56692 ± 0.0127107
高斯 + 选择性 k 近邻， $k = 5$	4.05446 ± 1.33153	15.4321 ± 4.98796	0.855387 ± 0.0261636	0.57206 ± 0.0129226
高斯 + 选择性 k 近邻， $k = 20$	4.05446 ± 1.33153	15.4321 ± 4.98796	0.862429 ± 0.0264602	0.56236 ± 0.0122017
高斯 + 选择性 k 近邻，L2范数	4.05446 ± 1.33153	15.4321 ± 4.98796	0.898932 ± 0.00921808	0.51486 ± 0.00474662

表 A33：Cells-Gap 数据集上的方法变体结果。

方法	验证集平均绝对误差 (↓)	验证集区间长度 (↓)	测试集覆盖率 (≥ 0.90)	测试集预测率（↑）
高斯 +选择性高斯混合模型	3.53089 ± 1.0619	15.6396 ± 6.23458	0.890569 ± 0.00953089	0.49372 ± 0.0025926
高斯 + 选择性高斯混合模型， $k = 2$	3.53089 ± 1.0619	15.6396 ± 6.23458	0.892265 ± 0.00867262	0.48902 ± 0.00420828
高斯 + 选择性高斯混合模型， $k = 8$	3.53089 ± 1.0619	15.6396 ± 6.23458	0.88967 ± 0.0111239	0.50196 ± 0.00492041
高斯 +选择性高斯混合模型，谱范数	2.81679 ± 0.42988	12.194 ± 1.10907	0.883644 ± 0.0167566	0.4991 ± 0.00761236
高斯 + 选择性 k 近邻	3.53089 ± 1.0619	15.6396 ± 6.23458	0.874032 ± 0.0432364	0.53646 ± 0.0139864
高斯 + 选择性 k 近邻， $k = 5$	3.53089 ± 1.0619	15.6396 ± 6.23458	0.874822 ± 0.0424219	0.5368 ± 0.0137332
高斯 + 选择性 k 近邻， $k = 20$	3.53089 ± 1.0619	15.6396 ± 6.23458	0.873876 ± 0.0443446	0.536 ± 0.014724
高斯 + 选择性 k 近邻，L2范数	3.53089 ± 1.0619	15.6396 ± 6.23458	0.887052 ± 0.0144573	0.49576 ± 0.00474578

表 A34：ChairAngle 数据集上的方法变体结果。

方法	验证集平均绝对误差 (↓)	验证集区间长度 (↓)	测试集覆盖率 (≥ 0.90)	测试集预测率 (↑)
高斯 +选择性高斯混合模型	0.376692 ± 0.171928	1.37757 ± 0.382191	0.902482 ± 0.00436054	0.972739 ± 0.00191733
高斯 + 选择性高斯混合模型， $k = 2$	0.376692 ± 0.171928	1.37757 ± 0.382191	0.902972 ± 0.00540708	0.96392 ± 0.00121759
高斯 + 选择性高斯混合模型， $k = 8$	0.376692 ± 0.171928	1.37757 ± 0.382191	0.901822 ± 0.00328732	0.981238 ± 0.00113892
高斯 +选择性高斯混合模型，谱范数	0.516487 ± 0.227015	1.82929 ± 0.809339	0.903828 ± 0.00534803	0.970993 ± 0.000466666
高斯 +选择性 k 近邻	0.376692 ± 0.171928	1.37757 ± 0.382191	0.90222 ± 0.00459694	0.975465 ± 0.00201785
高斯 + 选择性 k 近邻， $k = 5$	0.376692 ± 0.171928	1.37757 ± 0.382191	0.902561 ± 0.00342563	0.983804 ± 0.00152586
高斯 + 选择性 k 近邻， $k = 20$	0.376692 ± 0.171928	1.37757 ± 0.382191	0.902137 ± 0.00457917	0.966592 ± 0.00129222
高斯 +选择性 k 近邻，L2范数	0.376692 ± 0.171928	1.37757 ± 0.382191	0.903183 ± 0.00504343	0.977247 ± 0.00149241

表 A35：ChairAngle-Tails 数据集上的方法变体结果。

方法	验证集平均绝对误差 (↓)	验证集区间长度 (↓)	测试集覆盖率 (≥ 0.90)	测试集预测率 (↑)
高斯 +选择性高斯混合模型	0.241214 ± 0.091736	1.09417 ± 0.382907	0.901946 ± 0.00382993	0.655448 ± 0.001058
高斯 + 选择性高斯混合模型， $k = 2$	0.241214 ± 0.091736	1.09417 ± 0.382907	0.903353 ± 0.00471599	0.645452 ± 0.00155062
高斯 + 选择性高斯混合模型， $k = 8$	0.241214 ± 0.091736	1.09417 ± 0.382907	0.902103 ± 0.00382668	0.660472 ± 0.00115112
高斯 +选择性高斯混合模型, 谱范数	0.375118 ± 0.24509	1.24343 ± 0.611561	0.910147 ± 0.00315697	0.65518 ± 0.00176257
高斯 +选择性 k 近邻	0.241214 ± 0.091736	1.09417 ± 0.382907	0.860311 ± 0.00617031	0.703038 ± 0.00691227
高斯 + 选择性 k 近邻， $k = 5$	0.241214 ± 0.091736	1.09417 ± 0.382907	0.88625 ± 0.00498864	0.683742 ± 0.00547083
高斯 + 选择性 k 近邻， $k = 20$	0.241214 ± 0.091736	1.09417 ± 0.382907	0.809551 ± 0.00764886	0.743287 ± 0.0107676
高斯 +选择性 k 近邻，L2范数	0.241214 ± 0.091736	1.09417 ± 0.382907	0.89682 ± 0.00698641	0.662076 ± 0.00289246

表 A36：ChairAngle-Gap 数据集上的方法变体结果。

方法	验证集平均绝对误差 (↓)	验证集区间长度 (↓)	测试集覆盖率 (≥ 0.90)	测试集预测率 (↑)
高斯 +选择性高斯混合模型	0.454516 ± 0.280174	2.09212 ± 0.933756	0.91215 ± 0.00604745	0.649372 ± 0.00334919
高斯 + 选择性高斯混合模型， $k = 2$	0.454516 ± 0.280174	2.09212 ± 0.933756	0.912708 ± 0.00541793	0.642138 ± 0.00143869
高斯 + 选择性高斯混合模型， $k = 8$	0.454516 ± 0.280174	2.09212 ± 0.933756	0.911374 ± 0.00475979	0.654824 ± 0.00368979
高斯 +选择性高斯混合模型，谱范数	0.294511 ± 0.0861797	1.19904 ± 0.193646	0.910836 ± 0.00442911	0.649996 ± 0.00233428
高斯 +选择性 k 近邻	0.454516 ± 0.280174	2.09212 ± 0.933756	0.911574 ± 0.00376608	0.673764 ± 0.0113432
高斯 + 选择性 k 近邻， $k = 5$	0.454516 ± 0.280174	2.09212 ± 0.933756	0.91224 ± 0.00426519	0.670646 ± 0.00691764
高斯 + 选择性 k 近邻， $k = 20$	0.454516 ± 0.280174	2.09212 ± 0.933756	0.907543 ± 0.00204724	0.680107 ± 0.0162927
高斯 +选择性 k 近邻，L2范数	0.454516 ± 0.280174	2.09212 ± 0.933756	0.913383 ± 0.00539352	0.658744 ± 0.00274677

表 A37：AssetWealth 数据集上的方法变体结果。

方法	验证集平均绝对误差 (↓)	验证集区间长度 (↓)	测试集覆盖率 (≥ 0.90)	测试集预测率 (↑)
高斯 +选择性高斯混合模型	0.367501 ± 0.0416437	1.597 ± 0.207599	0.850824 ± 0.047533	0.93838 ± 0.0170443
高斯 + 选择性高斯混合模型， $k = 2$	0.367501 ± 0.0416437	1.597 ± 0.207599	0.850232 ± 0.0469534	0.933687 ± 0.0176348
高斯 + 选择性高斯混合模型， $k = 8$	0.367501 ± 0.0416437	1.597 ± 0.207599	0.85067 ± 0.0456954	0.932879 ± 0.0102523
高斯 +选择性高斯混合模型，谱范数	0.354647 ± 0.00973911	1.55594 ± 0.0421589	0.872742 ± 0.0250168	0.941812 ± 0.0101895
高斯 +选择性 k 近邻	0.367501 ± 0.0416437	1.597 ± 0.207599	0.852107 ± 0.0474734	0.933586 ± 0.0203541
高斯 + 选择性 k 近邻， $k = 5$	0.367501 ± 0.0416437	1.597 ± 0.207599	0.85256 ± 0.0469571	0.933232 ± 0.02149
高斯 + 选择性 k 近邻， $k = 20$	0.367501 ± 0.0416437	1.597 ± 0.207599	0.851914 ± 0.0460655	0.933636 ± 0.0200087
高斯 +选择性 k 近邻，L2范数	0.367501 ± 0.0416437	1.597 ± 0.207599	0.853365 ± 0.0410864	0.897149 ± 0.0162693

表 A38：VentricularVolume 数据集上的方法变体结果。

方法	验证集平均绝对误差 (↓)	验证集区间长度（↓）	测试集覆盖率（≥ 0.90）	测试集预测率（↑）
高斯 +选择性高斯混合模型	12.7238 ± 1.52197	51.566 ± 3.52739	0.752046 ± 0.0529087	0.707994 ± 0.0208741
高斯 + 选择性高斯混合模型， $k = 2$	12.7238 ± 1.52197	51.566 ± 3.52739	0.745271 ± 0.0582068	0.790752 ± 0.0307666
高斯 + 选择性高斯混合模型， $k = 8$	12.7238 ± 1.52197	51.566 ± 3.52739	0.747625 ± 0.0584117	0.656583 ± 0.0258853
高斯 +选择性高斯混合模型，谱范数	11.6311 ± 0.483357	45.5475 ± 0.747647	0.734907 ± 0.00853342	0.719279 ± 0.0147984
高斯 +选择性 k 近邻	12.7238 ± 1.52197	51.566 ± 3.52739	0.735105 ± 0.0612413	0.911599 ± 0.0447475
高斯 + 选择性 k 近邻， $k = 5$	12.7238 ± 1.52197	51.566 ± 3.52739	0.735857 ± 0.0604531	0.916614 ± 0.0419073
高斯 + 选择性 k 近邻， $k = 20$	12.7238 ± 1.52197	51.566 ± 3.52739	0.734752 ± 0.0616982	0.909404 ± 0.0469634
高斯 + 选择性 k 近邻，L2范数	12.7238 ± 1.52197	51.566 ± 3.52739	0.740812 ± 0.049802	0.740439 ± 0.0222073

表 A39：BrainTumourPixels 数据集上的方法变体结果。

方法	验证集平均绝对误差 (↓)	验证集区间长度 (↓)	测试集覆盖率 (≥ 0.90)	测试集预测率 (↑)
高斯 +选择性高斯混合模型	21.0625 ± 0.358012	93.6284 ± 2.29916	0.883515 ± 0.0138279	0.973576 ± 0.0187252
高斯 + 选择性高斯混合模型， $k = 2$	21.0625 ± 0.358012	93.6284 ± 2.29916	0.881874 ± 0.0118346	0.977863 ± 0.0145103
高斯 + 选择性高斯混合模型， $k = 8$	21.0625 ± 0.358012	93.6284 ± 2.29916	0.880823 ± 0.01273	0.973417 ± 0.01856
高斯 +选择性高斯混合模型，谱范数	22.0729 ± 1.25273	95.0605 ± 3.40124	0.890506 ± 0.0112498	0.9738 ± 0.00857087
高斯 +选择性 k 近邻	21.0625 ± 0.358012	93.6284 ± 2.29916	0.891264 ± 0.00734602	0.947185 ± 0.0178434
高斯 + 选择性 k 近邻， $k = 5$	21.0625 ± 0.358012	93.6284 ± 2.29916	0.891103 ± 0.00680399	0.949392 ± 0.0173229
高斯 + 选择性 k 近邻， $k = 20$	21.0625 ± 0.358012	93.6284 ± 2.29916	0.891759 ± 0.00804523	0.94453 ± 0.0188456
高斯 +选择性 k 近邻，L2范数	21.0625 ± 0.358012	93.6284 ± 2.29916	0.879639 ± 0.0084247	0.981862 ± 0.0068585

表 A40：SkinLesionPixels 数据集上的方法变体结果。

方法	验证集平均绝对误差 (↓)	验证集区间长度（↓）	测试集覆盖率（≥ 0.90）	测试集预测率（↑）
高斯 +选择性高斯混合模型	105.417 ± 1.15178	535.139 ± 215.446	0.821515 ± 0.0137705	0.849579 ± 0.0206181
高斯 + 选择性高斯混合模型， $k = 2$	105.417 ± 1.15178	535.139 ± 215.446	0.825054 ± 0.0122698	0.844799 ± 0.0191092
高斯 + 选择性高斯混合模型， $k = 8$	105.417 ± 1.15178	535.139 ± 215.446	0.815696 ± 0.010454	0.862683 ± 0.0259293
高斯 + 选择性高斯混合模型、谱范数	107.531 ± 2.11824	607.435 ± 407.758	0.821797 ± 0.0350093	0.837273 ± 0.0205022
高斯 + 选择性 k 近邻	105.417 ± 1.15178	535.139 ± 215.446	0.813027 ± 0.0306586	0.927047 ± 0.00830718
高斯 + 选择性 k 近邻， $k = 5$	105.417 ± 1.15178	535.139 ± 215.446	0.812763 ± 0.0309198	0.928375 ± 0.0100103
高斯 + 选择性 k 近邻， $k = 20$	105.417 ± 1.15178	535.139 ± 215.446	0.812419 ± 0.0306294	0.929172 ± 0.0109976
高斯 + 选择性 k 近邻、L2范数	105.417 ± 1.15178	535.139 ± 215.446	0.808667 ± 0.0112832	0.831784 ± 0.0181658

表 A41：HistologyNucleiPixels 数据集上的方法变体结果。

方法	验证集平均绝对误差 (↓)	验证集区间长度 (↓)	测试集覆盖率 (≥ 0.90)	测试集预测率 (↑)
高斯 +选择性高斯混合模型	211.795 ± 10.0239	1211.83 ± 396.946	0.59554 ± 0.0956742	0.90569 ± 0.0453555
高斯 + 选择性高斯混合模型， $k = 2$	211.795 ± 10.0239	1211.83 ± 396.946	0.597998 ± 0.094508	0.912836 ± 0.0323147
高斯 + 选择性高斯混合模型， $k = 8$	211.795 ± 10.0239	1211.83 ± 396.946	0.595718 ± 0.0973458	0.943449 ± 0.0409511
高斯 + 选择性高斯混合模型，谱范数	206.492 ± 7.82601	896.779 ± 61.4921	0.65672 ± 0.0824111	0.855668 ± 0.0632114
高斯 + 选择性 k 近邻	211.795 ± 10.0239	1211.83 ± 396.946	0.608929 ± 0.0805954	0.846228 ± 0.0482832
高斯 + 选择性 k 近邻， $k = 5$	211.795 ± 10.0239	1211.83 ± 396.946	0.610025 ± 0.0809455	0.842258 ± 0.046231
高斯 + 选择性 k 近邻， $k = 20$	211.795 ± 10.0239	1211.83 ± 396.946	0.607876 ± 0.0817183	0.858491 ± 0.0496696
高斯 + 选择性 k 近邻，L2范数	211.795 ± 10.0239	1211.83 ± 396.946	0.588514 ± 0.0914286	0.975562 ± 0.0190356

表 A42：AerialBuildingPixels 数据集上的方法变体结果。

方法	验证集平均绝对误差 (↓)	验证集区间长度（↓）	测试集覆盖率（≥ 0.90）	测试集预测率 (↑)
高斯 +选择性高斯混合模型	217.877 ± 1.72493	929.562 ± 47.6606	0.76535 ± 0.0388677	0.652082 ± 0.0990489
高斯 + 选择性高斯混合模型， $k = 2$	217.877 ± 1.72493	929.562 ± 47.6606	0.751721 ± 0.0454772	0.634602 ± 0.122618
高斯 + 选择性高斯混合模型， $k = 8$	217.877 ± 1.72493	929.562 ± 47.6606	0.850213 ± 0.025791	0.617738 ± 0.0879951
高斯 +选择性高斯混合模型，谱范数	220.763 ± 7.23119	1014.77 ± 164.155	0.817517 ± 0.0505544	0.638149 ± 0.046255
高斯 +选择性 k 近邻	217.877 ± 1.72493	929.562 ± 47.6606	0.651867 ± 0.0771154	0.840103 ± 0.0463989
高斯 + 选择性 k 近邻， $k = 5$	217.877 ± 1.72493	929.562 ± 47.6606	0.654256 ± 0.0771566	0.846067 ± 0.0456245
高斯 + 选择性 k 近邻， $k = 20$	217.877 ± 1.72493	929.562 ± 47.6606	0.647779 ± 0.0795474	0.832596 ± 0.0542738
高斯 +选择性 k 近邻，L2范数	217.877 ± 1.72493	929.562 ± 47.6606	0.815075 ± 0.0278439	0.641851 ± 0.11347