arXiv:2304.01426v1 [cs.LG] 4 Apr 2023

保形化无条件分位数回归

Ahmed M. Alaa
加州大学伯克利分校

Zeshan Hussain
MIT

David Sontag
MIT

摘要

我们提出了一种预测推断程序，将保形预测（CP）与无条件分位数回归（QR）相结合；后者是计量经济学中常用的工具 [1]，它涉及对输入协变量进行重心化影响函数（RIF）回归。与更广为人知的条件分位数回归不同，无条件分位数回归明确捕捉协变量分布变化对结果边际分布分位数的影响。利用这一特性，我们的方法能够生成具有局部频率学派覆盖率保证的自适应预测区间。该方法使用训练数据拟合 RIF 的机器学习模型，然后针对任意测试协变量，围绕新实例构造一个“假设的”协变量分布，并据此应用保形预测程序，使其局部化。实验表明，我们的方法能够适应异方差性，提供与当前测试实例相关且清晰透明的覆盖率保证，并且在效率方面与现有方法相比具有竞争力。

1 引言

考虑一个训练数据集 $(X_1, Y_1) \ldots, (X_n, Y_n)$，以及一个测试点 (X_{n+1}, Y_{n+1})，其中训练数据和测试数据均独立地从同一分布中抽取，即

$$(X_i, Y_i) \overset{i.i.d}{\sim} P = P_X \times P_{Y|X}, i = 1, \ldots, n+1. \quad (1)$$

这里，每个 $X_i \in \mathbb{R}^d$ 都是一个协变量向量，而 $Y_i \in \mathbb{R}$ 是一个响应变量。协变量与响应变量的联合分布 P未知。本文探讨预测推断问题：给定新测试点的协变量向量 X_{n+1}，目标是构造一个预测区间，使其以至少 $1-\alpha$ 的概率包含真实响应Y_{n+1}，其中 $\alpha \in$ 取值范围为 (0，1)。更确切地说，我们的目标是利用 n 个训练样本点构造一个集合值函数：

$$\widehat{C}_n(x) \coloneqq \widehat{C}_n((X_1, Y_1), \ldots, (X_n, Y_n), x) \subseteq \mathbb{R}, \quad (2)$$

使得对于一个新的测试点 $(X_{n+1}, Y_{n+1}) \sim P$，响应 Y_{n+1} 以概率 $1-\alpha$ 落在 $C_n(X_{n+1})$ 内。满足此覆盖率 条件的区间称为有效的。

预测推断有效的含义决定了相应覆盖率保证对特定预测实例的相关性。最弱形式的有效性是预测区间平均而言能够覆盖真实响应——这样的区间称为边际有效。形式上，当满足以下条件时，边际有效性成立：

$$\mathbb{P}\left[Y_{n+1} \in \widehat{C}_n(X_{n+1})\right] \geq 1 - \alpha, \quad (3)$$

其中，概率是针对训练数据和测试数据两者的随机性定义的。若（3）中的覆盖条件对所有 P都成立，则称其为分布无关的。保形预测（CP）在章节2中作了正式介绍，是一种广受欢迎的预测推断框架，能够在有限样本下保证分布无关的边际有效性 [2–8]。在最基本的形式中，CP 为所有预测实例给出一个固定长度的区间，从而实现边际有效性。

在许多应用中，确保在向个体用户传达预测结果的不确定性时保持透明度非常重要。例如，假设 X_i 是患者 i 的一组风险因素（如年龄、血压等），而 Y_i是肾功能的一项指标（如估算肾小球滤过率）。对于一位新患者，我们的目标是预测其未来估算肾小球滤过率可能达到的数值范围，并给出预先设定的置信度。也就是说，我们希望能够作出类似这样的陈述：“根据您的风险因素，未来 3 年内您的估算肾小球滤过率有 95% 的概率会下降 1.8–5.2 mL/min。”边际覆盖保证预测区间平均而言对 95% 的患者准确，但对特定预测实例的准确性可能任意低。由于边际覆盖率是相对于 P_X定义的，我们预计，在协变量空间中训练样本较少的区域，覆盖率会无法达到保证的水平；也就是说，对于这些实例，预测模型具有较高的认知不确定性。因此，

人工智能与统计国际会议（AISTATS）2023 年论文集，西班牙瓦伦西亚。PMLR：第 206 卷。版权所有 2023，作者所有。

由标准保形预测（CP）给出的具有边际有效性的固定长度区间，可能无法为那些最需要量化不确定性的预测实例提供有用信息。

自适应且透明的保形预测。本文探讨以下问题：我们如何才能以自适应且透明的方式，在具体的预测实例中传达模型不确定性？也就是说，我们希望构建一种预测推断程序，能够根据不同预测实例的不确定性水平，自适应调整其输出区间的长度，并报告与每个具体实例“相关”的覆盖率保证。理想情况下，我们希望开发一种能够实现以下条件覆盖率保证的预测推断程序：

$$\mathbb{P}\left[Y_{n+1} \in \widehat{C}_n(X_{n+1}) \,\middle|\, X_{n+1} = x\right] \geq 1-\alpha, \qquad (4)$$

对于几乎所有[1] $x \in \mathbb{R}^d$。满足（4）的预测区间称为条件有效。满足（4）的程序具有透明性，因为其保证适用于每个预测^实例；如果 C_n 对于给定的 x能够反映该具体实例的不确定性相对水平，则该程序具有自适应性。众所周知，对于非平凡预测，无法实现（4）意义下的分布无关有效性[6, 9, 10]。因此，我们基于保形预测框架开发一种自适应预测推断程序，使其满足（4）的近似形式，并透明地报告每个预测实例的覆盖率粒度。

对于每个新的测试点，我们的方法会报告一个针对该实例的预测区间，并确定一个包含该实例 $X_{n+1} = x$的局部区域；在该区域内，该方法具有边际有效性，即推断结果如下：

> ^ ^对于 $X_{n+1} = x$，报告 $C_n(x)$ $\mathcal{S}_n(x)$，使得：
>
> $$\mathbb{P}\left[Y \in \widehat{C}_n(X) \,\middle|\, X \in \widehat{\mathcal{S}}_n(x)\right] \geq 1-\alpha, \ \forall x \in \mathbb{R}^d,$$
>
> 其中 $\widehat{\mathcal{S}}_n(x) \subseteq \mathbb{R}^d$，我们称之为相关性子组，它是一个包含 x的局部区域。

在前文讨论的临床示例中，我们会这样向患者说明不确定性：“模型预测，您的估算肾小球滤过率将在未来 3 年内下降 1.8–5.2 mL/min。对于患者^与您相似、即被定义为患者子群 $\mathcal{S}_n(x)$的患者，模型的预测通常有 95% 的准确率，但该组内不同患者的预测准确率会有所不同。”通过以这种更透明的方式说明不确定性，临床医生可以查看所报告的子组，从而判断该预测与当前患者的相关性，例如，检查相关性子组是否包含具有不同疾病表型的患者。

[1]我们用“几乎所有 x”表示：在 P_X下，违反该界限的点集测度为零。

我们方法背后的关键思想是，将局部预测推断视为一个边际保形预测问题：针对“假设的”协变量分布 G_x（该分布以测试点x 为中心进行局部化），而不是针对真实分布 P_X。这一方法通过计量经济学中常用的一种方法实现，即无条件分位数回归（UQR）；该方法估计协变量空间 [1]任意局部子集（即相关性子组）中的结果边际分位数。具体而言，它对分位数泛函的重中心化影响函数（RIF）进行协变量回归，并在相关性子组内对预测得到的 RIF 求边际平均。这与条件分位数回归 [8], 不同；在条件分位数回归中，对回归目标在子组上取平均，无法得到边际分位数。

我们的方法分为两个步骤。首先，我们使用 UQR 模型为每个相关性子组生成一组嵌套的预测带。接下来，我们使用留出的校准集，选择每个子组中能够达到覆盖率的最窄预测带。在本文其余部分，我们将介绍这两个步骤；首先，我们会在下一章节简要介绍标准的保形预测方法。

2 保形预测

标准的拆分 保形预测（CP）流程依靠样本拆分来构造满足有限样本覆盖率保证的预测区间 [2–4]。假设所有数据点均可交换，该流程将数据集拆分为两个互不相交的子集：适当训练集 $\{(X_i, Y_i) : i \in \mathcal{D}_t\}$，以及校准 集 $\{(X_i, Y_i) : i \in \mathcal{D}_c\}$。随后，使用训练数据集 $\mathcal{D}_t$ 拟合机器学习模型 $\widehat{\mu}(x)$，并为 $\mathcal{D}_c$中的所有样本计算一致性分数 $V(。)$——该分数衡量相较于先前示例而言，预测结果看起来有多异常。 $V(。)$的一种常见选择是绝对残差，即 $V(x, y) := |\widehat{\mu}(x) - y|$。一致性分数的计算如下：

$$V_i := V(X_i, Y_i) = |\widehat{\mu}(X_i) - Y_i|, \forall i \in \mathcal{D}_c. \qquad (5)$$

对于给定的漏覆盖率 α，接下来计算绝对残差经验分布的一个分位数，

$$Q_{\mathcal{V}}(1-\alpha) := (1-\alpha)(1 + 1/|\mathcal{D}_c|)\text{-th quantile of } \mathcal{V}, \quad (6)$$

其中 $\mathcal{V} = \{V_i : i \in \mathcal{D}_c\}$。最后，新点 $X_{n+1} = x$ 处的预测区间为

$$\widehat{C}_n(x) = [\widehat{\mu}(x) - Q_{\mathcal{V}}(1-\alpha), \widehat{\mu}(x) + Q_{\mathcal{V}}(1-\alpha)]. \quad (7)$$

保形预测区间的长度固定为 $2Q_{\mathcal{V}}(1-\alpha)$，不依赖于 X_{n+1}。这足以满足边际有效性，但无法适应不同预测实例的不确定性程度差异。

3 保形化无条件分位数回归（CUQR）

在本章节中，我们介绍该方法所涉及的两个步骤，并将其称为保形化无条件分位数回归（CUQR）。事实上，我们并非首个提出保形预测自适应变体的方法；我们将在第4节中比较我们的方法与现有的自适应不确定性量化方法。

3.1 步骤 1：无条件分位数回归（UQR）

考虑 C_n 在（4）中的真实总体对应形式，即

$$C^*(x) := [Q(\alpha/2, x), Q(1-\alpha/2, x)], \quad (8)$$

其中，C^* 是基于预言机已知 P 时的预测区间，$Q(\alpha, x)$ 是 $Y|X=x$ 的 α 水平分位数，即 $Q(\alpha, x) := \inf\{y \in \mathbb{R} : F(y|X=x) \geq \alpha\}$，而 $F(.)$ 是条件累积分布函数（CDF），$F(y|X=x) := \mathbb{P}(Y \leq y|X=x)$。根据定义，（8）中的预言机不确定性带在（4）的意义下是条件有效的。因此，可以通过直接估计条件分位数 $Q(.,x)$，得到一个兼具自适应性和透明性的不确定性带的合理猜想。

插件估计的嵌套序列。

我们采用插件方法来估计 C^*，方法是将式（8）中的条件分位数替换为一致估计量 Q，即 $\widehat{C}_n(x) = [\widehat{Q}(\alpha/2, x), \widehat{Q}(1-\alpha/2, x)]$。虽然插件模型能够准确估计 C^*，但无法提供有限样本覆盖保证。为了兼顾插件估计的自适应性与保形预测框架的有限样本覆盖，典型做法是对这些插件估计进行“保形化” [8]。接下来，我们将说明如何根据 C^* 的插件估计构造一致性分数。

我们不使用单个插件模型 $\widehat{Q}(\alpha,.)$ 得到的 $Q(\alpha,.)$ 点估计来构造预测区间，而是生成一组条件分位数函数的“候选”估计，并使用校准集 $\mathcal{D}_c$ 选出能够达到所需覆盖率的最窄候选带。更准确地说，我们为协变量 x, $\mathcal{C}(x)$ 定义一组预测区间，如下所示：

$$\widehat{\mathcal{C}}(x) := \{\widehat{\mathcal{C}}_{\tilde{\alpha}}(x) = \widehat{\mu}(x) \pm \widehat{Q}(\tilde{\alpha}, x)\}_{\tilde{\alpha} \in (0,1)}, \quad (9)$$

其中，$\widehat{Q}(\tilde{\alpha}, x)$ 是在 x 处模型残差的第 $\tilde{\alpha}$ 级条件分位数的插件估计。我们要求这些插件估计具有单调性：$\widehat{Q}(\tilde{\alpha}, x) \leq \widehat{Q}(\tilde{\alpha}', x)$ 对于 $\tilde{\alpha} \leq \tilde{\alpha}'$，即不存在分位数交叉。因此，$\mathcal{C}(x)$ 构成了一个嵌套的候选区间序列，我们据此定义以下一致性分数：

$$V(X_i, Y_i) = \inf\{\tilde{\alpha} \in (0,1) : Y_i \in \widehat{\mathcal{C}}_{\tilde{\alpha}}(X_i)\}, \quad (10)$$

对于所有 $i \in \mathcal{D}_c$。式（10）中的一致性分数会检验 $\tilde{\alpha}$ 的最小取值，使得相应区间 $\widehat{\mathcal{C}}_{\tilde{\alpha}}(X_i)$ 在 $\mathcal{C}$ 中能够覆盖响应变量 Y_i。我们按照式（$Q_{\mathcal{V}}(1-\alpha)$ 6）计算一致性分数的经验分位数，并为 $X_{n+1} = x$ 构造预测区间，如下所示：

$$\widehat{C}_n(x) := \{\widehat{\mu}(x) \pm \widehat{Q}(\tilde{\alpha}^*, x)\}, \; \tilde{\alpha}^* = Q_{\mathcal{V}}(1-\alpha). \quad (11)$$

为简化记号，我们省略了 C_n 对 α 的依赖。请注意，式（11）中的方法从嵌套集合中选取整个条件分位数函数，因此生成的预测区间会随预测实例而变化。式（11）中的区间仍遵循保形预测构造，因此满足以下边际覆盖保证。

命题 1. 考虑一列插件估计 $\{\widehat{Q}(\tilde{\alpha}, .)\}_{\tilde{\alpha}}$，它们由样本 $\mathcal{D}_c$,*1* 得到；相应的一致性分数 $\mathcal{V} = \{V(X_i, Y_i) : i \in \mathcal{D}_c,2\}$ 则由另一个样本 $\mathcal{D}_c$,*2* 得到，其中 $\mathcal{D}_c$,*1* 和 $\mathcal{D}_c$,*2* 是 $\mathcal{D}_c$ 的两个不相交子集。若 $\{(X_i, Y_i) : 1 \leq i \leq n+1\}$ 可交换，则式（*11*）中的区间满足

$$\mathbb{P}(Y_{n+1} \in \widehat{C}_n(X_{n+1})) \geq 1 - \alpha.$$

证明见附录 A。文献中也有这一结果的变体 [7, 11]。命题 1 表明，通过在一列带而非区间上定义一致性分数，我们可以构造自适应预测区间，同时保留保形预测的边际覆盖保证。

通过无条件分位数回归进行插件估计。

我们使用无条件分位数回归 [1] 来拟合式（9）中的嵌套插件估计序列。下文将从泰勒近似的角度解释无条件分位数回归的工作原理。无条件分位数回归将 $Y|X=x$，$Q(\alpha, x)$ 在真实分布 P 下的条件分位数，近似为 Y，$Q(\alpha)$ 在另一种分布下的边际分位数 G_x，该分布围绕 x 进行“局部化”。由于分位数是底层分布的统计泛函，我们可以利用冯·米塞斯线性近似（VOM）估计 G_x 下的边际分位数，已知 P 下的边际分位数。该近似是泰勒级数的一种分布类比，形式如下 [12]：

$$Q_{G_x}(\alpha) \approx Q_P(\alpha) + \int \text{IF}(y; Q(\alpha), P) \cdot dG_x(y), \quad (12)$$

其中，IF 是泛函 $Q(\alpha)$ 在 P 处、沿局部化分布 G_x 方向的影响函数；这里给定的点为 (x, y)，其定义如下：

$$\text{IF}(y; Q(\alpha), P) = \lim_{\epsilon \to 0} \frac{Q_{P^y}(\alpha) - Q_P(\alpha)}{\epsilon}, \quad (13)$$

其中 $P^y = (1-\epsilon)P + \epsilon \delta_y$。分位数的影响函数衡量结果值 y 对边际分位数统计量 $Q_P(\alpha)$ 的贡献。按照式（12）使用局部化密度 P 对从 G_x 中抽样的观测值进行加权后，我们得到一阶

近似，即在 G_x 下对边际分位数泛函的近似。分位数泛函的影响函数为：

$$\mathrm{IF}(y;Q(\alpha),P)=\frac{\alpha-\mathbf{1}\{y\le Q_P(\alpha)\}}{f_Y(Q_P(\alpha))}. \tag{14}$$

这里，$f_Y(.)$ 是Y的边际密度。(14) 中公式的推导是标准的，为完整起见，附录 B 中给出了推导过程。重新中心化影响函数（RIF）的定义如下：

$$\mathrm{RIF}(y;Q(\alpha),P)=Q_P(\alpha)+\mathrm{IF}(y;Q(\alpha),P), \tag{15}$$

无条件分位数回归是对 X 进行回归，以得到关于$\mathbb{E}[\mathrm{RIF}(Y;\ Q(\alpha)\ P)$ 的$|\,X=x\,]$。请注意，(14) 中的影响函数是一个二元变量，因为 $\mathbf{1}\{y\le Q_P(\alpha)\}$ 是唯一一个随协变量变化的项。因此，无条件分位数回归需要为 f_Y拟合一维密度估计，并为该二元变量拟合一个二元分类器。无条件分位数回归通常用于研究改变协变量分布对结果边际分位数的影响，例如工会化对工资的影响 [13]。在我们的设定中，我们使用(12) 在测试点处获得预测带的插件估计 x，如下所示：

$$Q_{G_x}(\alpha)\approx\int\mathbb{E}[\,\mathrm{RIF}(Y;Q(\alpha),P)\,|\,X=x\,]\cdot dG_x. \tag{16}$$

使用无条件分位数回归构造嵌套序列。

式（16）中的近似启发了一种简单的回归方法，可在避免分位数交叉的同时构造式（9）中的嵌套序列。令 $\mathrm{RIF}_\alpha(X_i)$ 表示与第 i 个数据点对应的 α 级分位数的 RIF。对于一个测试点 $X_{n+1}=x$，我们可以通过在数据集 $\{(X_i,\mathrm{RIF}_\alpha(X_i))\}_{i\in\mathcal{D}_c}$ 上拟合一个机器学习模型，来预测 $Q_{G_x}(\alpha)$ 的值。$\widehat{\ }$令机器学习模型预测的 RIF 值为 $\mathrm{RIF}_\alpha(x)$。然后，对于 $\tilde\alpha=[1/K,\dots,(K-1)/K]$ 中的所有 $\tilde\alpha$ 值，重复此过程 $K>0$ 次，即可构造如下嵌套序列：[2]

$$\widehat{\mathcal{C}}(x)=\{\widehat{\mathcal{C}}_k(x)=\widehat{\mu}(x)\pm|\widehat{\mathrm{RIF}}_{\tilde\alpha_k}(x)|\}_{k=1}^{K-1}. \tag{17}$$

其中 $\tilde\alpha_k=1/k$。这里，RIF 是针对模型残差 $E=|\widehat{\mu}(X)-Y|$ 的分位数定义的，而非针对结果变量 Y。注意，结合式（14）和式（15），级分位数的 RIF 可写为：

$$\mathrm{RIF}_{\tilde\alpha_k}(x)=Q_P(\tilde\alpha_k)+\frac{\tilde\alpha_k-\mathbf{1}\{e\le Q_P(\tilde\alpha_k)\}}{f_E(Q_P(\tilde\alpha_k))}. \tag{18}$$

一维密度 f_E(。) 可以使用核密度估计（KDE）进行估计，而 $Q_P(\alpha)$ 可以估计为残差的经验水平-$\tilde\alpha_k$ 分位数。对于每个测试点，我们唯一需要预测的项是 $\mathbf{1}\{e\le Q_P(\tilde\alpha_k)\}$，$\forall k\in\{1,\dots,K-1\}$。具体而言，可以通过以下方式使用单个机器学习模型实现：对于每个数据点 (X_i,E_i)，定义目标 $k_i^*:=\min\ k,\ \text{s.t.}\ E_i\le Q_P(\tilde\alpha_k)$，然后在数据集 $\{(X_i,k_i^*)\}_i$ 上拟合模型 g_θ(.)（例如回归模型或多分类器）。对于新的测试点 $X_{n+1}=x$，将 g_θ 的预测值代入（18），即可预测 $X_{n+1}=x$ 处的 RIF：

$$\widehat{\mathrm{RIF}}_{\tilde\alpha_k}(x)=\widehat{Q}_P(\tilde\alpha_k)+\frac{\tilde\alpha_k-\mathbf{1}\{g_\theta(x)\le k\}}{\widehat{f}_E(\widehat{Q}_P(\tilde\alpha_k))}, \tag{19}$$

$\forall k\in\{1,\dots,K-1\}$。因此，可以使用 (19) 中的 $K-1$ 预测值来构造 (17) 中的嵌套集合。请注意，对于任意 $k<k'$，只要满足 $\partial\widehat{f}_E(\widehat{F}_E^{-1}(\alpha))/\partial\alpha<0$，就足以保持嵌套区间的单调性，$\widehat{\ }$即 $\widehat{\mathrm{RIF}}^{\tilde\alpha_k}(x)<\widehat{\mathrm{RIF}}^{\tilde\alpha_{k'}}(x)$。多种典型的概率分布（例如指数分布）都满足此条件。如果在我们的经验估计中，任意两个连续区间不满足此条件，我们就用 $f_E(Q_P(\tilde\alpha_{k-1}))$ 替换 $\widehat{f}_E(Q_P(\tilde\alpha_k))$，以确保单调性。

$\widehat{\ }$构造集合 $\{Q(\alpha,.)\}_\alpha$ 的其他方法包括拟合 K 独立的分位数回归模型，或单个分布模型（例如贝叶斯非参数回归 [14]）。基于 RIF 的构造所得到的嵌套$\widehat{\ }$集合 $\{Q(\alpha,.)\}_\alpha$ 在计算和统计上都比这两种方法更高效，因为它只需训练一个机器学习模型，其标签汇总了关于 $[1/K,\dots,(K-1)/K]$各水平条件分位数的信息。

3.2 步骤 2：在子组内对无条件分位数回归进行保形化

在章节 3.1中，我们提出了一种程序，既满足自适应性要求，又保留非自适应保形预测方法的边际有效性（命题 1）。然而，这些边际保证并不够透明，因为它们无法反映针对任何给定预测所输出区间的准确性。为了提供更透明的保证，第二步通过在协变量空间的局部区域内应用保形程序，从嵌套集合中选取一个区间，具体如下：

- 将协变量空间划分为 G 个子集 $\{\widehat{\mathcal{S}}_g\}_{g=1}^G$。
- 在每个子组 g 内，局部应用章节3.1 中的保形程序，方法是从集合 $\{\widehat{\mathcal{C}}_k(x)\}_{k=1}^{K-1}$ 中选取一个子组特定带 $\widehat{\mathcal{C}}_{k\ (g)}(x)$。
- 对于测试点 $X_{n+1}=x$，确定其所属的相关子组 g_{n+1}，并报告该子组 $\widehat{\mathcal{S}}_{g_{n+1}}$ 及相应的区间 $\widehat{\mathcal{C}}_{k\ (g_{n+1})}(X_{n+1})$。

我们的保形化 UQR（CUQR）流程所涉及的步骤见算法 1。该流程既会报告针对每个实例的预测区间，也会报告一个能够达到所需准确率水平的子组。相关性子组既可以从训练

[2]我们取 $\widehat{\ }$RIF 的绝对值，以避免出现错误的负预测值。

算法 1：保形化 UQR

输入： 数据集 $\{(X_i, Y_i)\}_{i=1}^n$，(X_{n+1}, Y_{n+1}) α，μ，G，K；测试协变量 X_{n+1}，模型 μ，参数 α、G 和 K。

输出： $\widehat{C}_n(X_{n+1})$ 和 $\widehat{\mathcal{S}}_n(X_{n+1})$

将数据集划分为适当训练集 $\mathcal{D}_t$ 和两个不相交的校准集 $\mathcal{D}_c = \mathcal{D}_{c_1} \cup \mathcal{D}_{c_2}$；使用训练集 $\mathcal{D}_t$ 执行以下操作：；拟合预测模型 $\widehat{\mu}: \mathcal{X} \to \mathbb{R}$，使用 $\mathcal{D}_t$；将 $\mathcal{X}$ 划分为相关性子组 $\{\widehat{\mathcal{S}}_g\}_{g=1}^G$；使用校准集 $\mathcal{D}_c$，执行以下操作：；拟合 $\widehat{f}_E(.)$、$\widehat{Q}_P(\tilde{\alpha}_k)$ 和 g_θ。将估计值代入式（19）；构造区间的嵌套序列 $\widehat{\mathcal{C}}(x) = \{\widehat{\mathcal{C}}_k(x) = \widehat{\mu}(x) \pm |\widehat{\text{RIF}}^{\tilde{\alpha}_k}(x)|\}_{k=1}^K$；使用校准集 $\mathcal{D}_c$，执行以下操作：；对 $\{\widehat{\mathcal{S}}_g\}_{g=1}^G$ 中每个子组内的所有校准数据，计算式（10）中的一致性分数；对于每个子组，在 $\mathcal{C}$ 中选择达到目标覆盖率的最窄区间，如式（11）所示。令 $\mathcal{C}_{k(g)}(x)$ 表示为第 g 个子组选择的区间；

对于新的测试协变量 X_{n+1}，执行以下操作：；

确定满足条件的子组 g_{n+1} $X_{n+1} \in \mathcal{S}_{g_{n+1}}$；

$\widehat{C}_n(X_{n+1}) \leftarrow \mathcal{C}_{k(g_{n+1})}(X_{n+1})$，$\widehat{\mathcal{S}}_n(X_{n+1}) \leftarrow \mathcal{S}_{g_{n+1}}$；

返回 $\widehat{C}_n(X_{n+1})$ 和 $\widehat{\mathcal{S}}_n(X_{n+1})$。

数据（例如，使用聚类算法）或根据特定应用领域的知识预先确定（例如，疾病亚型或受保护属性）。该子组可以表示为一个簇，并附上一个对于该子组具有代表性的协变量值。在前文讨论的临床示例中，我们的算法可以用一组具有代表性的患者来表示相关性子组。

增加子群数量 G 会提高所达到的平均覆盖率的粒度，但代价是实际覆盖率的方差增大。对于较大的样本量，可以将 G 设为较大的值，例如 $G = O(n)$，从而渐近地实现条件覆盖率。在另一极端情况下，当 $G = 1$ 时，我们得到标准的边际覆盖率保证。接下来，我们陈述 CUQR 区间 $\widehat{C}_n(\cdot)$ 所提供的理论覆盖率保证。

定理 1。 若 $\{(X_i, Y_i) : i \in \mathcal{D}_c\} \cup \{(X_{n+1}, Y_{n+1})\}$ 在给定 $\{\mathcal{S}_g\}_{g=1}^G$ 的条件下可交换，且 $\mathbb{P}(\mathcal{S}_g) > \delta$、$\forall g$ 和某些 $\delta > 0$，那么对于任意 $\alpha \in (0,1)$ 和 $n \in \mathbb{Z}_+$，我们有

$$\mathbb{P}\left[Y_{n+1} \in \widehat{C}_n(X_{n+1}) \,\middle|\, X_{n+1} \in \widehat{\mathcal{S}}_g, \mathcal{D}\right] \geq 1 - \alpha - \frac{\lambda}{\sqrt{n_g}},$$

$\forall 1 \leq g \leq G$，其概率至少为 $1 - 2\exp(-2\lambda^2)$，对任意 $\lambda \geq \sqrt{\log(2)/2}$均成立。定理 1 指出，对于每个相关子组 g，覆盖率以高概率成立，误差范围为 $\lambda/\sqrt{n_g}$，其中 n_g 是 S_g中的校准点数量。因此，通过选择经验覆盖率为 $(1-\alpha) + \lambda/\sqrt{n_g}$的预测区间，可以在每个子组内实现 PAC 覆盖率保证。这表明，输出区间的**透明度**与**效率**之间存在权衡：增加子群数量 G 意味着覆盖率几乎必然适用于输入空间中更细粒度的子集，即提高透明度。代价是区间会更长，因为随着 G 增大，n_g 会变小。定理 1 中的界由对符合度分数 [15]的经验累积分布函数应用 Dvoretzky-Kiefer-Wolfowitz 不等式得到。完整证明见附录 C。

4 相关工作

分布回归模型直接估计条件密度 $P_{Y|X=x}$，从而自适应地估计不确定性。许多方法都属于这一类，以下仅列举部分方法：贝叶斯非参数回归（例如使用高斯过程 [16]或回归树 [14]）、贝叶斯神经网络 [17–19], 和深度集成 [20, 21]。这些模型中有许多能够对预测方差提供准确且实用的估计，但无法实现保形预测（CP）所带来的有限样本覆盖保证。分布回归的实际覆盖率可能对建模选择（例如超参数或先验分布）非常敏感。例如，在贝叶斯回归中，采用精确推断时，后验可信区间所达到的频率学派覆盖率取决于先验分布的选择 [22]。而使用更常见的近似推断方法（例如随机失活或变分推断 [23]）时，所诱导的后验分布可能不会渐近集中，导致覆盖表现较差 [24, 25]。分布模型常与共形预测方法结合使用，以满足有限样本覆盖保证，同时保持自适应性 [26]。

保形预测 以牺牲自适应性为代价，实现了最基本形式下的有限样本边际覆盖率[2–4]。近期提出了多种基于保形预测的自适应预测推断方法 [7, 8, 10, 11, 27–31]。“保形化”条件分位数函数的插件估计这一思想最早源于 [8]。在这项工作中，使用一个分位数回归模型 $Q(\alpha, x)$ 对训练数据进行拟合，并使用衡量 $Q(\alpha, x)$ 准确性的 一致性分数来为这些区间推导调整量。保形化分位数回归提供了

边际覆盖保证，但其经验条件覆盖率取决于底层分位数回归模型的质量。随后，人们沿着两条不同的研究路线对这一方法进行了改进：第一条路线使用重新加权技术，在新的测试点处对保形预测进行“局部化”[27, 30], ；第二条路线则对分布回归模型进行保形化，并可从中推导出条件分位数 [26, 29]。在这两条研究路线中，条件有效性均是在渐近意义下实现的。

我们的工作与这两种方法存在微妙的联系。在一致性分数的构造方面， [26, 29]采用条件秩而非误差残差来定义分数——我们的方法通过从改变分位数水平 α所生成的条件分位数“候选”估计中进行选择，来构造预测区间。 $P_{Y|X}$这等同于从完整条件分布的条件秩估计中进行选择来构造预测区间， [26, 29], 但与 $P_{Y|X}$中的方法不同，我们的方法不需要获取[27, 30], 的一致估计。与[27, 30], 中的局部化保形预测方法类似，我们的方法实质上是边际保形预测方法的局部化版本。但与局部化保形预测（LCP）不同，我们的方法可以使用任何机器学习模型来获得局部分位数估计（即公式（19））， 而 LCP 则仅限于重新加权估计器（例如基于 Nadaraya–Watson 核的估计器）。此外，由于我们的方法在每个子组内从多个分位数函数中进行选择，因此能够在局部区域内实现有限样本（而非渐近）覆盖率。我们的插件估计的嵌套构造属于 [11]提出的一般嵌套保形预测框架。与该工作中的框架不同，我们通过参数化预测带的函数形式来构造嵌套集合，而不是直接参数化区间；这样便能为子组中的每个实例选择不同的区间。

重新中心化影响函数（RIF）通常用作无条件分位数回归模型的目标，这是计量经济学研究中常用的一种建模工具 [1]。据我们所知，在本研究之前，基于 RIF 的回归尚未被用作一致性分数。

5 实验

我们在多个基准数据集上，将 CUQR 与各种保形方法和分位数回归基线方法进行比较。下面先介绍我们的实验设置。

基线方法。 我们考虑标准的分割式保形预测（CP）、局部自适应保形预测 （LACP）方法 [32], 保形化分位数回归（*CQR*） [8], ，以及保形条件直方图（CCH） [29]。我们还考虑两种标准分位数回归（QR）模型的变体，用于估计条件分位数：一种基于随机森林模型的 QR（QR-RF），另一种使用神经网络实现的 QR（QR-NN）。我们考虑 CUQR 的两个消融版本

它们在 CUQR 的相同相关性子组内应用保形化过程。第一个基线方法称为 CQ，它利用相关性子组内残差的经验分位数构造嵌套集合 $\mathcal{C}$（。），并将相同的区间分配给子组内所有样本单位，不使用基于 UQR 的插件估计。第二个基线方法称为 CQR-S，它在相关性子组内应用自适应 CQR 方法 [8]。

最后，我们考察所提方法的两个变体：CUQR 基于经验 $(1-\alpha)$ 分位数进行保形化；CUQR-PAC 则对定理 1 中的松弛项进行校正，从而为每个子组提供高概率（PAC 风格）的覆盖率。我们选择定理 1 中 λ 的取值，使得覆盖率成立的概率（以训练数据为条件）为 90%，即 $1-2\exp(-2\lambda^2) = 0.9$。我们使用 XGBoost 回归模型来实现针对 g_θ 的方法。在所有实验中，我们使用 K-均值聚类算法来构建相关性子群。更多实验细节见附录 D。

评估指标。 我们根据所有基线方法达到的边际覆盖率 C_{av}、通过平均区间长度 L_{av} 量化的效率，以及记作 $C_{av}(\widehat{\mathcal{S}}_g)$ 的所有子组 $\{\widehat{\mathcal{S}}_g\}_{g=1}^{G}$的子组层面覆盖率来评估这些方法。我们还评估最坏情况子组级覆盖率，将其定义为 $C_G^{w.c.} = \min_g C_{av}(\widehat{S}_g)$。所有指标均在测试数据上评估，并取 10 次运行的平均值。除非另有说明，我们将目标覆盖率设为 $1-\alpha = 0.9$。

数据集。 我们在**9 个基准数据集** 上评估所有基线方法，这些数据集常用于评估保形预测方法：MEPS-19、MEPS-20、MEPS-21、Facebook-1、Facebook-2、Bio、Kin8nm、Naval 和 Blog [8, 28, 28,33]。受篇幅限制，本章节重点展示四个数据集（MEPS-19、Facebook-1、Blog 和 Kin8nm）的结果，其余结果见附录。所有数据集的详细信息均见附录。每次运行时，我们都将每个数据集随机划分为互不重叠的训练集 $\mathcal{D}_t$（42.5%）、校准集 $\mathcal{D}_c$（42.5%）和测试集 $\mathcal{D}_{test}$（15%）。对于 LACP、CCH、CQR、CQR-S、CQ 和 CUQR 基线方法，我们还将校准集 $\mathcal{D}_c$ 对半划分，以便从不同的划分中获得插件估计和一致性分数。在所有实验中，我们使用训练集 $\mathcal{D}_t$ 拟合梯度提升回归模型 $\widehat{\mu}$，并基于该模型给出的预测值 $\widehat{\mu}$应用预测推断基线方法。

结果

评估透明度。 所有基线方法均经过校准，使其目标边际覆盖率达到 90%，但对于模型的个体用户而言，这种覆盖率概念意味着什么？在本实验中，我们通过评估典型个体或预测实例“子组”中的最坏情况条件覆盖率，来评估不同基线方法的透明度。我们使用 K-均值聚类，利用训练样本识别 $G = 10$ 相关性子组

	MEPS-19			Facebook-1			Blog			Kin8nm		
	C_{av}	L_{av}	$C_G^{w.c.}$	C_{av}	L_{av}	$C_G^{w.c.}$	C_{av}	L_{av}	$C_G^{w.c.}$	C_{av}	L_{av}	$C_G^{w.c.}$
QR方法												
QR-RF	0.90	1.00	0.54	0.93	0.85	0.78	0.79	0.73	0.76	0.93	1.36	0.89
QR-NN	0.79	0.54	0.67	0.81	0.55	0.68	0.79	0.73	0.76	0.79	0.94	0.74
CP方法												
CP	0.89	1.28	0.19	0.90	1.39	0.72	0.89	1.89	0.57	0.90	2.17	0.83
LACP	0.89	0.61	0.20	0.90	0.69	0.76	0.89	1.06	0.63	0.90	1.09	0.84
CQR	0.89	1.12	0.46	0.90	0.83	0.77	0.90	1.34	0.82	0.90	1.33	0.85
CCH	0.96	5.37	0.79	0.89	0.72	0.65	0.98	5.58	0.96	0.89	1.14	0.86
CQ	0.87	2.02	0.76	0.89	1.34	0.79	0.87	1.81	0.76	0.89	2.16	0.85
CQR-S	0.89	1.54	0.67	0.90	0.77	0.87	0.90	1.37	0.80	0.90	1.33	0.85
CUQR	0.89	1.25	0.73	0.90	1.36	0.87	0.87	1.82	0.67	0.89	2.19	0.85
CUQR-PAC	0.89	2.90	0.88	0.92	1.61	0.90	0.90	2.23	0.82	0.96	3.27	0.93

表1：基准数据集上所有基线方法的边际覆盖率、效率和条件覆盖率。

。请注意，这些子组并非任意划分的，它们代表了将总体聚类为由相似个体组成的“典型”子组。

表1展示了三个数据集上所有基线方法的预测区间边际覆盖率和平均长度。首先，我们观察到，虽然所有保形方法都达到了目标（边际）覆盖率，但QR基线方法的覆盖率会因底层模型设定而异，这正如预期，因为这些基线方法不提供任何覆盖率保证。相反，所有基于CP的方法都实现了其承诺的模型无关覆盖率保证，但在子组层级上考察时，不同CP变体的表现如何？

如表1所示，仅保证边际覆盖率的保形预测基线方法（保形预测和CQR）在子组条件下的最坏情况覆盖率非常低；也就是说，它们所宣称的保证无法反映其在给定总体中大量“相似”个体中的表现。类似地，具有渐近条件覆盖率保证的保形预测方法，在有限样本中可能出现严重的覆盖不足（LACP），也可能维持合理的条件覆盖率，但效率较低（CCH）。这凸显了控制有限样本条件覆盖率的重要性。虽然CUQR能够在边际意义上实现子组特定的覆盖率，但对于某些数据集（例如MEPS-19和Blog），最坏情况子组级覆盖率可能显著低于期望的目标覆盖率。CUQR-PAC

我们方法的变体以高概率保证每个子组内的覆盖率。这是以效率为代价的，即预测区间会更宽。

我们方法的共形分位数变体为每个子组选择一个固定区间，而不是完整的基于RIF的预测带。它以牺牲子群内自适应性和平均效率为代价，实现了更好的最坏情况覆盖率。Be-

由于 CUQR 的子组特定覆盖率保证是相对于校准数据的随机性平均成立的，因此随着子组数量 G 增加（即每个子组的校准样本更少），测试数据上经验覆盖率的方差也会增大。因此，随着相关子组数量增加，CUQR 所达到的最差子组层面覆盖率会（在期望意义上）降低。为权衡透明度与效率，我们方法中的 CUQR-PAC 变体通过扩大预测区间来考虑实际覆盖率的方差，从而达到经验覆盖率 $(1-\alpha)+\lambda/\sqrt{n_g}$（见图 S1 1）。

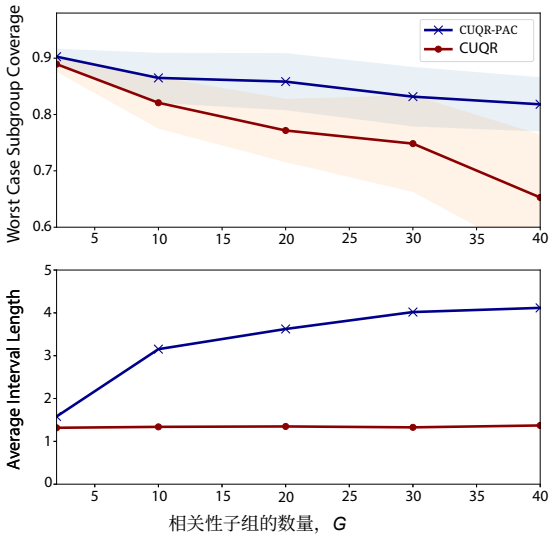


图1：子群数量 G 对覆盖率的影响。

评估自适应性。接下来，我们评估

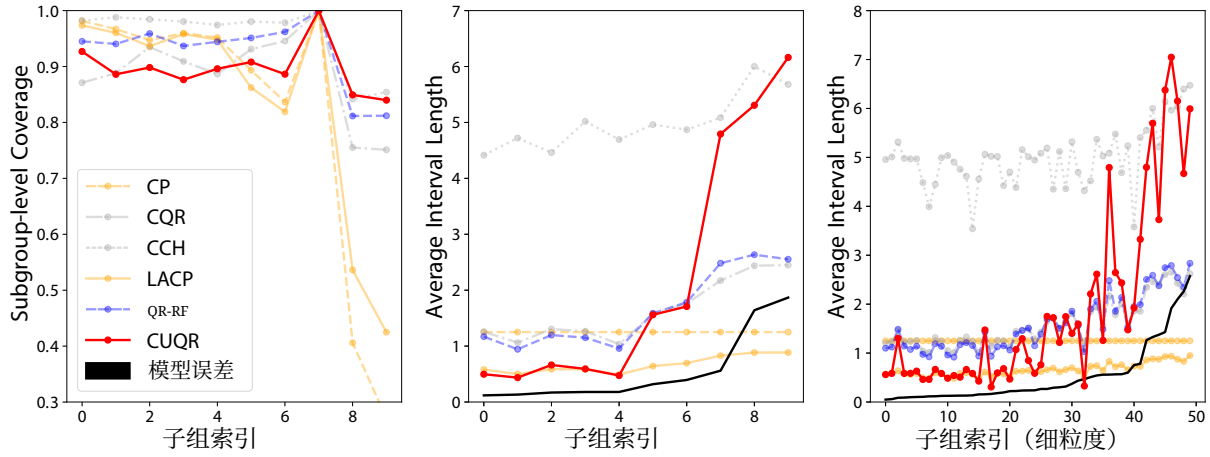


图2：**MEPS-19** 数据集上预测推断基线方法的自适应性。

基线方法是自适应的，即其区间长度会随基础模型的真实不确定性而变化 $\hat{\mu}$。在所研究的数据集中，MEPS-19 数据集表现出显著的异方差性，即预测模型的平均误差在不同子组之间有显著差异。因此，我们预计，对于模型误差较高的子组，自适应方法输出的区间平均长度会更大。在图 2中，我们绘制了 MEPS-19 数据集的子组层面实际覆盖率（左）以及每个子组对应的平均区间长度（中）。在这两幅图中，**MEPS-19**数据集中的子组索引沿 x 轴按测试数据上的模型子组平均误差从低到高排列（即索引越大，模型不确定性越高）。（在图 2中，为避免图示过于拥挤，我们省略了表现较差的基线方法。）可以看出，CUQR 对所有子组都能大致维持目标覆盖率，并根据模型不确定性调整各子组内输出区间的长度。相反，其他基线方法要么无法识别不同子组间不确定性的差异（LACP 和 CP），要么无法充分调整区间长度以维持目标覆盖率（QR-RF 和 CQR）。

最后，为了评估 CUQR 在用于校准的子组之外的自适应性，我们使用不同的随机种子和不同的簇数量运行 K均值聚类算法 $G = 50$，并像之前一样，根据基础模型的不确定性按升序排列子组索引。我们在新的、更细粒度的 50 个子组上评估 CUQR（此前已针对 $G = 10$ 个子组进行拟合）以及其他基线方法的平均区间长度。如图2 （右）所示，CUQR 在根据子组层面的不确定性调整区间方面优于其他基线方法，这表明其条件自适应性优于理论保证所能体现的水平。

6 结论

本文提出了一种保形预测方法，该方法会根据每个预测实例的不确定性水平调整其输出区间，同时报告协变量空间中的一个局部区域。该区域包含所查询的协变量实例，并且该方法保证在该区域内的平均准确率。我们的方法将协变量空间划分为若干子组，并利用分位数泛函的重新中心化影响函数构造一个嵌套的预测带序列，再为每个子组选择一个预测带。通过向最终用户报告实例特定的预测区间和子组特定的覆盖率保证，我们的方法能够以更透明的方式传达机器学习模型预测中的不确定性。

参考文献

[1] SergioFirpo、Nicole M Fortin 和 ThomasLemieux。无条件分位数回归。《计量经济学》，77(3):953–973，2009年。

[2] Volodya Vovk、Alexander Gammerman 和 Craig Saunders。算法随机性的机器学习应用。1999年。

[3] Vladimir Vovk、Alexander Gammerman 和 Glenn Shafer。随机世界中的算法学习。SpringerScience& Business Media，2005年。

[4] Harris Papadopoulos。 归纳式保形预测：理论及其在神经网络中的应用。 INTECH Open Access Publisher Rijeka，2008年。

[5]Jing Lei、James Robins 和 Larry Wasserman。无分布预测集。美国统计学会杂志，108(501):278–287，2013年。

[6] Jing Lei 和 Larry Wasserman。非参数回归的分布无关预测带。英国皇家统计学会杂志：*B*辑：统计方法学，第 71–96 页，2014年。

[7] Jing Lei、Max G'Sell、Alessandro Rinaldo、Ryan J Tib-shirani 和 Larry Wasserman。用于回归的分布无关预测推断。《美国统计学会杂志》，113(523):1094–1111，2018年。

[8] 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，32:3543–3553，2019年。

[9] Vladimir Vovk. 归纳式保形预测器的条件有效性。载于亚洲机器学习会议，第475–490页。PMLR，2012年。[10] Rina Foygel Barber、Emmanuel J Candes、Aaditya Ramdas 和 Ryan J Tibshirani。分布无关条件预测推断的局限性。信息与推断：*IMA*期刊，10(2):455–482，2021年。[11] Chirag Gupta、Arun K Kuchibhotla 和 Aaditya Ramdas。嵌套保形预测与分位数袋外集成方法。模式识别，第108496页，2021年。[12] Luisa Turrin Fernholz。统计泛函的冯·米塞斯演算，第19卷。Springer Science & Business Media，2012年。[13] Stephen R Porter。分位数回归：分析分布变化，而非均值变化。高等教育：理论与研究手册，第335–381页，2015年。[14] Hugh A Chipman、Ed ward I George 和 Robert E McCulloch。贝叶斯加性回归树：贝叶斯加性回归树。应用统计年刊，4(1):266–298，2010年。[15] Pascal Massart。Dvoretzky-Kiefer-Wolfowitz不等式中的紧常数。概率年刊，第1269–1283页，1990年。[16] Suzanne Sniekers 和 Aad van der Vaart。高斯过程先验下回归问题的自适应贝叶斯可信集。电子统计学期刊，9(2):2475–2527，2015年。[17] Durk P Kingma、Tim Salimans 和 Max Welling。变分随机失活与局部重参数化技巧。神经信息处理系统进展，28，2015年。

[18] José Miguel Hernández-Lobato 和 Ryan Adams。用于可扩展贝叶斯神经网络学习的概率反向传播。载于国际机器学习会议，第1861–1869页。PMLR，2015年。

[19] Hippolyt Ritter、Aleksandar Botev 和 David Barber。用于神经网络的可扩展拉普拉斯近似。载于第六届国际表征学习会议 *ICLR 2018* 会议论文集，第6卷。国际表征学习会议，2018年。

[20]Balaji Lakshminarayanan、Alexander Pritzel 和 Charles Blundell。使用深度集成进行简单且可扩展的预测不确定性估计。神经信息处理系统进展，30，2017年。

[21]Stanislav Fort、Huiyi Hu 和 Balaji Lakshmi-narayanan。深度集成：从损失景观的角度出发。*arXiv* 预印本 *arXiv:1912.02757*，2019年。[22] M Jésus Bayarri 和 James O Berger。贝叶斯分析与频率学派分析的相互作用。统计科学，19(1):58–80，2004年。[23] Yarin Gal 和 Zoubin Ghahramani。随机失活作为贝叶斯近似：在深度学习中表示模型不确定性。载于国际机器学习会议，第1050–1059页。PMLR，2016年。[24] Ian Osband。深度学习中的风险与不确定性：贝叶斯、自助法及随机失活的危险。载于*NIPS* 贝叶斯深度学习研讨会，第192卷，2016年。[25] Jiri Hron、Alexander G de G Matthews 和 Zoubin Ghahramani。变分高斯随机失活并非贝叶斯方法。*arXiv* 预印本 *arXiv:1711.02989*，2017年。[26] Victor Chernozhukov、Kaspar Wüthrich 和 Yinchu Zhu。分布式保形预测。美国国家科学院院刊，118(48)，2021年。[27] Leying Guan。通过局部化实现保形预测。*arXiv* 预印本 *arXiv:1908.08558*，2019年。[28] Shai Feldman、Stephen Bates 和 Yaniv Romano。通过正交分位数回归提升条件覆盖率。神经信息处理系统进展，34，2021年。[29] Matteo Sesia 和 Yaniv Romano。使用条件直方图进行保形预测。神经信息处理系统进展，34，2021年。[30] Leying Guan。局部化保形预测：一种用于保形预测的广义推断框架。*arXiv* 预印本 *arXiv:2106.08460*，2021年。

[31] Yu Bai、Song Mei、Huan Wang、Yingbo Zhou 和 Caiming Xiong。基于一般函数类的高效且可微分保形预测。*arXiv* 预印本 *arXiv:2202.11091*，2022年。

[32] Harris Papadopoulos、Vladimir Vovk 和 Alexander Gammerman。基于最近邻的回归保序预测。《人工智能研究杂志》，40:815–840，2011。

[33] Youngseog Chung、Willie Neiswanger、Ian Char 和 Jeff Schneider。《超越 pinball 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，34，2021年。

补充材料

附录 A：命题 1 的证明

^命题 *1*。考虑从校准样本 $\mathcal{D}_c$、*1* 中获得的一系列插件估计 $\{Q(\tilde{\alpha}, .)\}_{\tilde{\alpha}}$，以及从另一个样本 $\mathcal{D}_c$、*2* 中获得的相应符合度分数 $\mathcal{V} = \{V(X_i, Y_i) : i \in \mathcal{D}_{c,2}\}$，其中 $\mathcal{D}_c$、*1* 和 $\mathcal{D}_c$、*2* 是 $\mathcal{D}_c$ 的两个不相交子集。若 $\{(X_i, Y_i) : 1 \leq i \leq n+1\}$ 是可交换的，则式（*11*）中的区间满足

$$\mathbb{P}(Y_{n+1} \in \widehat{C}_n(X_{n+1})) \geq 1 - \alpha.$$

证明。 区间在 (11) 中的构造意味着，在给定$\mathcal{D}_{c,1}$的条件下，我们有

$$Y_{n+1} \in \widehat{C}_n(X_{n+1}) \iff V(X_{n+1}, Y_{n+1}) \leq Q_{\mathcal{V}}(1-\alpha),$$

^因此可得 $\mathbb{P}(Y_{n+1} \in C_n(X_{n+1}) | \mathcal{D}_{c,1})) = \mathbb{P}(V(X_{n+1}, Y_{n+1} \leq Q_{\mathcal{V}}(1-\alpha) | \mathcal{D}_{c,1}))$。由于所有训练样本和校准样本 (X_i, Y_i) 都是可交换的，因此在给定 $\mathcal{D}_{c,1}$的条件下， $\mathcal{V}$ 中的校准一致性分数与测试点 $V(X_{n+1}, Y_{n+1})$ 上的一致性分数是可交换的。根据校准样本与测试样本的一致性分数的可交换性，由 [8] 中的引理 2 可得：

$$\mathbb{P}(V(X_{n+1}, Y_{n+1}) \leq Q_{\mathcal{V}}(1-\alpha) \,|\, \mathcal{D}_{c,1}) \geq 1 - \alpha,$$

对$\mathcal{D}_{c,1}$的随机性取边缘化后，即可得出该结论。 □

附录 B：分位数泛函的影响函数

回忆一下， $Q(\alpha, x)$ 是 $Y|X = x$ 的 α 级分位数，即：

$$Q(\alpha, x) = F^{-1}(\alpha) := \inf\{y \in \mathbb{R} : F(y \,|\, X = x) \geq \alpha\},$$

其中， $F(.)$ 是条件累积分布函数（CDF）， $F(y \,|\, X = x) := \mathbb{P}(Y \leq y \,|\, X = x)$。在本章节中，我们假设 $\alpha = F(F^{-1}(\alpha))$。回忆一下，对于给定点 (x, y)，泛函 $Q(\alpha)$ 在分布 P 处沿局部化分布 G_x 方向的影响函数定义如下：

$$\text{IF}(y; Q(\alpha), P) = \lim_{\varepsilon \to 0} \frac{Q_{P^y}(\alpha) - Q_P(\alpha)}{\varepsilon}, \tag{20}$$

其中 $P_\varepsilon = (1-\varepsilon)P + \epsilon\, \delta_y$。设 $F_\varepsilon = (1-\varepsilon)F + \varepsilon \cdot \delta_y$ 为扰动分布的累积分布函数，则有

$$\begin{aligned} \alpha &= F_\varepsilon \circ F_\varepsilon^{-1}(\alpha) \\ &= (1-\varepsilon) \cdot F(F_\varepsilon^{-1}(\alpha)) + \varepsilon \cdot \delta_y(F_\varepsilon^{-1}(\alpha)). \end{aligned} \tag{21}$$

对等式两边关于 ε 求导，我们得到如下结果

$$\begin{aligned} \frac{\partial \alpha}{\partial \varepsilon} = 0 &= \frac{\partial}{\partial \varepsilon} \left(F_\varepsilon \circ F_\varepsilon^{-1}(\alpha)\right) \\ &= \frac{\partial}{\partial \varepsilon} \left((1-\varepsilon) \cdot F(F_\varepsilon^{-1}(\alpha)) + \varepsilon \cdot \delta_y(F_\varepsilon^{-1}(\alpha))\right) \\ &= -F(F_\varepsilon^{-1}(\alpha)) + (1-\varepsilon) f(F_\varepsilon^{-1}(\alpha)) \cdot \frac{\partial F_\varepsilon^{-1}(\alpha)}{\partial \varepsilon} + \delta_y(F_\varepsilon^{-1}(\alpha)) + \varepsilon \cdot \frac{\partial}{\partial \varepsilon} \delta_y(F_\varepsilon^{-1}(\alpha)). \end{aligned} \tag{22}$$

注意，(20) 中影响函数的定义可以改写如下：

$$\text{IF}(y; Q(\alpha), P) = \lim_{\varepsilon \to 0} \frac{Q_{P^y}(\alpha) - Q_P(\alpha)}{\varepsilon} = \left(\frac{\partial}{\partial \varepsilon} F_\varepsilon^{-1}(\alpha)\right)_{\varepsilon=0}. \tag{23}$$

令 $\varepsilon = 0$代入（22），可得：

$$\begin{aligned} 0 &= \left(-F(F_\varepsilon^{-1}(\alpha)) + (1-\varepsilon) f(F_\varepsilon^{-1}(\alpha)) \cdot \frac{\partial F_\varepsilon^{-1}(\alpha)}{\partial \varepsilon} + \delta_y(F_\varepsilon^{-1}(\alpha)) + \varepsilon \cdot \frac{\partial}{\partial \varepsilon} \delta_y(F_\varepsilon^{-1}(\alpha))\right)_{\varepsilon=0} \\ &= -F(F^{-1}(\alpha)) + f(F^{-1}(\alpha)) \cdot \left(\frac{\partial F_\varepsilon^{-1}(\alpha)}{\partial \varepsilon}\right)_{\varepsilon=0} + \delta_y(F^{-1}(\alpha)). \end{aligned} \tag{24}$$

重新整理（24）中的各项，可得

$$\left(\frac{\partial F_{\varepsilon}^{-1}(\alpha)}{\partial \varepsilon}\right)_{\varepsilon=0} = \frac{F(F^{-1}(\alpha)) - \delta_y(F^{-1}(\alpha))}{f(F^{-1}(\alpha))}. \tag{25}$$

因此，分位数的影响函数为：

$$\mathrm{IF}(y; Q(\alpha), P) = \left(\frac{\partial F_{\varepsilon}^{-1}(\alpha)}{\partial \varepsilon}\right)_{\varepsilon=0} = \frac{\alpha - \mathbf{1}_{\{F^{-1}(\alpha)\geq y\}}}{f(F^{-1}(\alpha))}. \tag{26}$$

附录 C：定理 1 的证明

我们首先陈述将用于证明定理 1 的 Dvoretzky-Kiefer-Wolfowitz（DKW）不等式。设 F_n 为具有分布函数 F 的 n 个独立同分布随机变量的经验分布函数，则有

$$\mathbb{P}\left(\sup_x (\widehat{F}_n(x) - F(x)) > \frac{\lambda}{\sqrt{n}}\right) \leq \exp(-2\lambda^2), \tag{27}$$

对于任意满足 $\exp(-2\lambda^2) \leq \frac{1}{2}$ 的 λ，或等价地，$\lambda \geq \sqrt{\log(2)/2}$。证明见 [15]。

定理 1。如果 $\{(X_i, Y_i)\colon\ i \in \mathcal{D}_c\} \cup \{(X_{n+1}, Y_{n+1})\}$ 在给定 $\{\mathcal{S}_g\}_{g=1}^{G}$的条件下是可交换的，并且 $\mathbb{P}(\mathcal{S}_g) > \delta, \forall g$ 以及某个 $\delta > 0$，那么对于任意 $\alpha \in (0,1)$ 和 $n \in \mathbb{Z}_+$，我们有

$$\mathbb{P}\left[Y_{n+1} \in \widehat{C}_n(X_{n+1}) \,\middle|\, X_{n+1} \in \widehat{\mathcal{S}}_g, \mathcal{D}\right] \geq 1 - \alpha - \frac{\lambda}{\sqrt{n_g}},$$

$\forall 1 \leq g \leq G$，以至少 $1 - 2\exp(-2\lambda^2)$ 的概率，对任意 $\lambda \geq \sqrt{\log(2)/2}$。

证明。固定一个相关的 子组 $g \in \{1, \ldots, G\}$。通过在 11) 中构造区间，且取值于 $\mathcal{S}_g$，可得覆盖事件：

$$Y_{n+1} \in \widehat{C}_n(X_{n+1}) \,|\, X_{n+1} \in \widehat{\mathcal{S}}_g \iff V(X_{n+1}, Y_{n+1}) \leq Q_{\mathcal{V}_g}(1-\alpha) \,|\, X_{n+1} \in \widehat{\mathcal{S}}_g, \tag{28}$$

其中 $\mathcal{V}_g = \{V(X_i,\ Y_i)\colon\ i \in \mathcal{D}_c,\ X_i \in \widehat{S}_g\}$。由于在给定 $\{\widehat{S}_{g'}\}_{g'=1}^{G}$ 的条件下，$\{(X_i,\ Y_i)\colon i \in \mathcal{D}_c\}$ 可交换，因此一致性分数序列 $\mathcal{V}_g$ 和 $V(X_{n+1},\ Y_{n+1})$ 在给定 $\{\widehat{S}_{g'}\}_{g'=1}^{G}$ 及事件 $X_{n+1} \in \widehat{\mathcal{S}}_g$ 的条件下可交换。因此，由 [8] 中的引理 2 可得：

$$\mathbb{P}(V(X_{n+1}, Y_{n+1}) \leq Q_{\mathcal{V}_g}(1-\alpha) \,|\, X_{n+1} \in \widehat{\mathcal{S}}_g) \geq 1 - \alpha.$$

给定训练数据 $\mathcal{D}$，我们可以通过对符合度分数的累积分布函数经验估计应用 DKW 不等式，得到覆盖事件概率的上界如下：

$$\mathbb{P}\left(\sup_v \left(\frac{1}{n_g}\sum_{i=1}^{n_g} \mathbf{1}_{\{V(X_{n+1},Y_{n+1})\leq v\}} - F(v)\right) > \frac{\lambda}{\sqrt{n_g}} \,\middle|\, X_{n+1} \in \widehat{\mathcal{S}}_g, \mathcal{D}\right) \leq \exp(-2\lambda^2), \tag{29}$$

因此可得

$$\mathbb{P}\left(\frac{1}{n_g}\sum_{i=1}^{n_g} \mathbf{1}_{\{V(X_{n+1},Y_{n+1})\leq Q_{\mathcal{V}_g}(1-\alpha)\}} - (1-\alpha) > \frac{\lambda}{\sqrt{n_g}} \,\middle|\, X_{n+1} \in \widehat{\mathcal{S}}_g, \mathcal{D}\right) \leq \exp(-2\lambda^2), \tag{30}$$

$\forall \lambda \geq \sqrt{\log(2)/2}$。根据等价关系式 (28)，可以将式 (30) 中的上界写为：

$$\mathbb{P}\left(\frac{1}{n_g}\sum_{i=1}^{n_g} \mathbf{1}_{\{Y_{n+1}\in \widehat{C}_n(X_{n+1})\}} > 1 - \alpha - \frac{\lambda}{\sqrt{n_g}} \,\middle|\, X_{n+1} \in \widehat{\mathcal{S}}_g, \mathcal{D}\right) \leq \exp(-2\lambda^2), \tag{31}$$

由此得证。 □

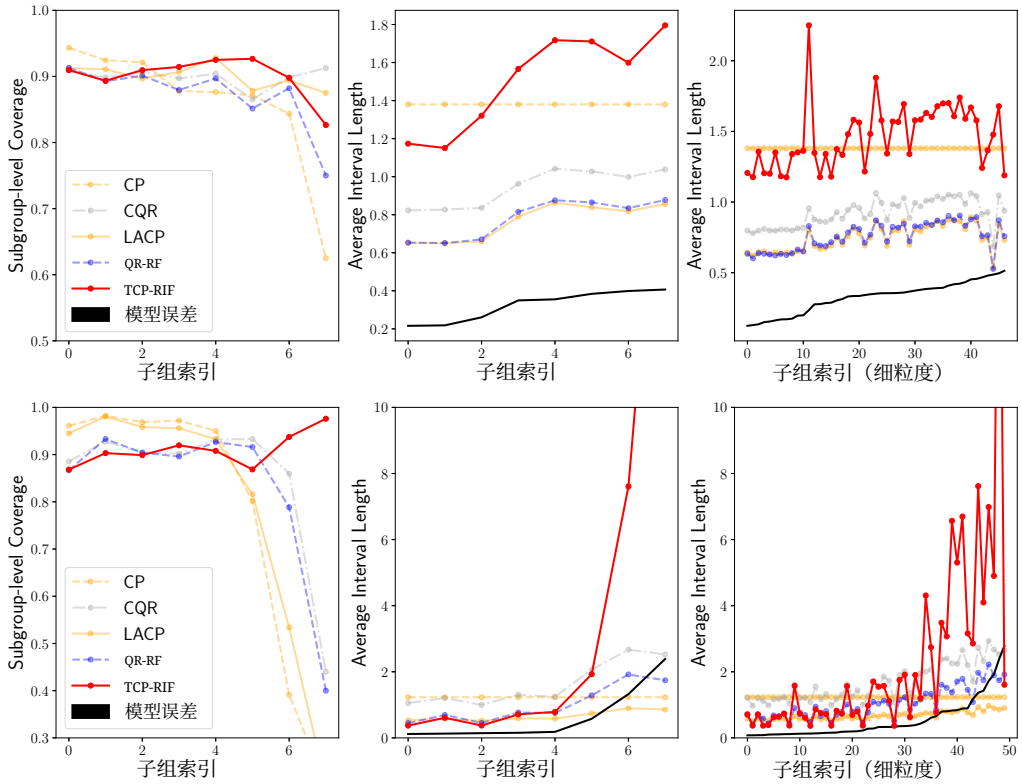


图 3：预测推断基线方法在**Facebook-1** 和 **MEPS-20** 数据集上的自适应性。

附录 D：其他实验

我们在 9 个数据集上测试了我们的方法：**MEPS-19**、**MEPS-20**、**MEPS-21**、**Facebook-1**、**Facebook-2**、**Bio**、**Kin8nm**、**Naval** 和**Blog**。MEPS（医疗支出面板调查）数据集包含对美国各地家庭和个人、其医疗服务提供者及雇主进行的调查，调查持续了三年 [8, 28, 33]。Facebook 数据集包含从 Facebook 帖子中提取的特征，任务是预测帖子将收到多少条评论。**Facebook-1**和**Facebook-2** 之间的区别在于所包含的特征。Bio 数据集由描述蛋白质三级结构理化性质的特征组成，包括蛋白质不同区域的分数面积、特定氨基酸之间的距离以及其他空间分布约束。任务是预测残基的大小。其余数据集可在 UCI仓库中获取。

评估透明度 在表 2中，我们展示了所有基线方法以及 CUQR 方法在另外三个数据集上的边际覆盖率、最坏情况子组覆盖率和平均区间长度。总体而言，我们再次观察到了正文中讨论的现象，不过我们会对一些更具竞争力的基线方法进行进一步分析。对于 QR-RF 基线方法，我们发现，尽管它在效率方面表现更优，但在边际覆盖率或条件覆盖率方面没有理论保证。事实上，在这三个数据集中，其条件覆盖率均低于 CUQR 所达到的最坏情况条件覆盖率。至于 CQR，尽管它在Facebook-2 数据集上表现颇具竞争力，但它在子组层级的覆盖率方面没有任何理论保证（仅保证边际覆盖率）。因此，我们看到Facebook-1 和 **Bio** 上的条件覆盖率不如 CUQR。

评估自适应性 与正文中所述的实验类似，我们继续评估基线方法的自适应程度，即区间长度与基础模型的真实不确定性（由模型误差反映）的一致程度。我们在**Facebook-1** 和**MEPS-20** 数据集上进行这项实验。和之前一样，我们发现，基线方法要么不具备自适应性（例如保形预测或 LACP），要么自适应性有限，不足以 在子组中维持目标覆盖率（例如 QR-RF、CQR）。具体而言，QR-RF 和保形预测在模型不确定性较高的子组（即索引较大的子组）中，覆盖率会大幅下降。另一方面，我们看到 CUQR 能够在所有子组中维持覆盖率，并根据子组的不确定性调整区间长度。

	MEPS-19			MEPS-20			MEPS-21		
	C_{av}	L_{av}	$C_G^{w.c.}$	C_{av}	L_{av}	$C_G^{w.c.}$	C_{av}	L_{av}	$C_G^{w.c.}$
QR方法									
QR-RF	0.90	1.00	0.54	0.91	1.06	0.55	0.93	1.29	0.65
QR-NN	0.79	0.54	0.67	0.81	0.56	0.64	0.79	0.53	0.64
CP方法									
CP	0.89	1.28	0.19	0.90	1.24	0.15	0.90	1.29	0.16
LACP	0.89	0.61	0.20	0.89	0.59	0.20	0.90	0.62	0.26
CQR	0.89	1.12	0.46	0.89	1.05	0.50	0.89	1.27	0.62
CCH	0.96	5.37	0.79	0.97	5.35	0.75	0.97	5.35	0.78
CQ	0.87	2.02	0.76	0.88	2.01	0.87	0.88	2.10	0.88
CUQR	0.89	1.25	0.73	0.89	1.21	0.76	0.90	1.26	0.81
	Facebook-1			**Facebook-2**			**Bio**		
	C_{av}	L_{av}	$C_G^{w.c.}$	C_{av}	L_{av}	$C_G^{w.c.}$	C_{av}	L_{av}	$C_G^{w.c.}$
QR方法									
QR-RF	0.93	0.85	0.78	0.92	0.81	0.86	0.92	1.33	0.84
QR-NN	0.81	0.55	0.68	0.80	0.52	0.64	0.81	0.94	0.56
CP方法									
CP	0.90	1.39	0.72	0.90	1.39	0.82	0.90	2.42	0.85
LACP	0.90	0.69	0.76	0.90	0.69	0.84	0.90	1.15	0.83
CQR	0.90	0.83	0.77	0.90	0.83	0.87	0.90	1.36	0.83
CCH	0.89	0.72	0.65	0.89	0.64	0.67	0.90	1.07	0.85
CQ	0.89	1.34	0.79	0.89	1.35	0.87	0.90	2.41	0.80
CUQR	0.90	1.36	0.87	0.89	1.36	0.87	0.90	2.40	0.88
	Kin8nm			**Naval**			**Blog**		
	C_{av}	L_{av}	$C_G^{w.c.}$	C_{av}	L_{av}	$C_G^{w.c.}$	C_{av}	L_{av}	$C_G^{w.c.}$
QR方法									
QR-RF	0.93	1.36	0.89	0.90	0.63	0.87	0.79	0.73	0.76
QR-NN	0.79	0.94	0.74	0.78	0.55	0.73	0.79	0.73	0.76
CP方法									
CP	0.90	2.17	0.83	0.89	1.31	0.78	0.89	1.89	0.57
LACP	0.90	1.09	0.84	0.89	0.60	0.85	0.89	1.06	0.63
CQR	0.90	1.33	0.85	0.89	0.70	0.85	0.90	1.34	0.82
CCH	0.89	1.14	0.86	0.89	0.48	0.83	0.98	5.58	0.96
CQ	0.89	2.16	0.85	0.87	1.27	0.87	0.87	1.81	0.76
CUQR	0.89	2.19	0.85	0.86	1.26	0.85	0.87	1.82	0.67

表 2：基准数据集上所有基线方法的边际覆盖率、效率和条件覆盖率

ts.