个性化保形推断

Fernando Delbianco - Fernando Tohm´e

南方国立大学经济系，巴伊亚布兰卡数学研究所，南方国立大学 - *CONICET*

arXiv:2304.05189v1 [stat.ME] 11 Apr 2023

摘要

可以使用保形预测的变体来解决个性化预测问题，从而获得感兴趣变量的实际值所属的区间。我们在此介绍一种方法：先检测可能与特定问题相关的观测值，再使用模拟对照组得出预测值的区间。结果表明，该方法具有自适应性，并且能够检测潜在相关变量是否存在。

关键词：保形预测、个体化推断、拆分与刀切分布无关推断。

1 引言

我们提出了一种新颖的个性化保形预测方法，旨在给出区间，以涵盖某些变量的实际预测值。我们首先采用一种类似于相关观测值的技术，针对给定的目标查询进行识别，这种技术借鉴了分而治之(Chen 等人 (2021））。然后，我们应用一种数据增强程序，类似于重复样本(Xie 和 Wang(2022））， 模拟对照组，仿佛它们是通过自助法生成的（如Tran 等人 (2017）。该方法基于一种称为保形预测的统计技术。

保形预测（CP）是一种通常基于机器学习系统的无分布假设预测方法。CP 基于带标签的训练数据集，对新的测试数据点生成预测，唯一假设是数据的可交换性。它要求预先指定一个水平，以限制算法允许犯错的频率（例如，参见 Vovk 等人 （2022）、Xie和 Zheng （2022）、Lei 等人 （2018）、Barber 等人 （2021））。该方法的变体包括拆分和杰克刀法无分布假设的保形推断（Lei 等人 （2018））。与我们的贡献密切相关的一项最新发展是 CP 的自适应版本，其基于自监督学习的应用

(Seedat 等人(2023)).

文献中与我们方法相关的另一条研究脉络是个体化推断。这是一种在大型数据集情况下特别有用的方法，因为它处理了数据异质性。一种方法基于生成与目标个体共享共同特征的数据$iGroups$ （Cai 等（2021））。另一种方法是，在$iFusion$ 方法中，相关数据组通过融合与目标个体相似的个体所作的推断来生成（Shen 等 （2020））。个体化推断还被用于通过混合由具有相关特征的个体生成的高斯过程来对案例进行分类（Alaa 和 Van Der Schaar （2017））。

用于进行推断的具体技术多种多样。 Chernozhukov 等人 (2017) 开发了一种去偏机器学习方法，用于对特定参数进行推断。也可以采用更传统的方法，例如推断 置信度 (Xie 和 Singh(2013), Schweder 和 Hjort (2016)) 以及 预测 (Shen 等人(2018)) 分布。这些方法可以视为“类贝叶斯频率学派技术”，它们根据观测数据推导出分布。在实践中，一项重要贡献是以 **R语言** 实现可训练的 p-值 函数 (而非仅仅是 p-值），由 Infanger 和 Schmidt-Trucks¨ass (2019)。

我们的方法具有以下特征：

- 它具有自适应性。
- 它接受新的无标签数据作为输入。
- 通过将关注范围限定于相关数据及其模拟数据，它满足可交换性条件。
- 它通过隐式消除导致独立同分布条件失效的因素，经验性地检测潜在 相关变量的存在。

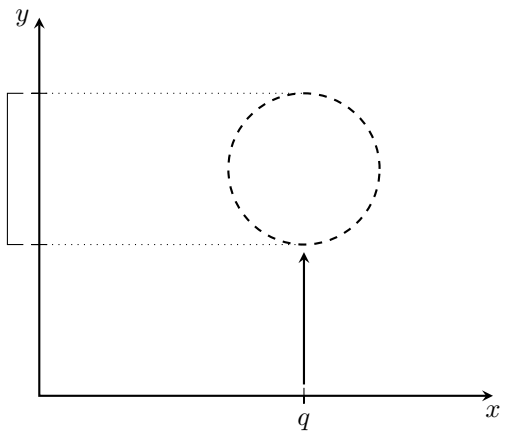
本文结构如下。首先，我们在第 2 节介绍提出本方法的动机。第 3 节详细说明实现所开发方法的流程。第 4 节介绍用于检验该方法的实证设置及相应结果。第 5 节总结全文，并提出进一步发展的思路。

2 动机

我们假设，如Delbianco 等人(2021所述，存在一个关于数据生成过程的统计模型，可描述为 $\{\mathcal{O}, \mathcal{P}\}$，其中 $\mathcal{O}$ 是观测值的集合，而 $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ 是定义在 $\mathcal{O}$ 上的一族概率分布，且 Θ 是模型的参数空间。目标仍然是针对任何查询q 估计参数的区间，其中 q 是在给定推断方法下对信息的特定请求 $\mathcal{I}$，该方法应用于数据库。

查询定义了多个维度。首先， $\mathcal{O}$ 由条目$\{o_i = \langle x_i, y_i \rangle\}_{i=1,\ldots,n}$组成，其中 x_i 是一个包含 p 个变量和 n 个观测值的矩阵，是 o_i的尾部，而 y_i 是其头部，是一个维度为 n的向量。查询 q由一个尾部 x_0组成，没有头部，并且对于 $i = 1$、. . .、 n，具有 $x_o \neq x_i$。我们假设存在一类潜变量 $\mathcal{S}$，使得给定一个查询 q，便存在相应的 $s_q \in \mathcal{S}$，从而得到一类相关观测 $\mathcal{O}_q \subseteq \mathcal{O}$ [1]。

图 1：潜在相关区域 S_q



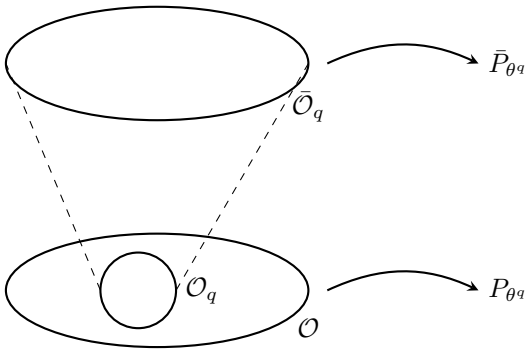
尾部 q的某个特定值与潜变量 s_q 相关联，该潜变量会生成尾部的相关空间$x \in X_q$，并因此能够推断出相应的头部类别 $y \in Y_q$。接着，可以将这个二维图扩展为三维图，并将不同查询的序列作为第三个轴。这表示相关集合如何随 q变化，并学习相应的 θ^q。

然后，我们生成一类对照组 $\mathcal{O}_q$，以验证 $\mathcal{O}_q \subseteq \mathcal{O}_q$。基于 $\mathcal{O}_q$ 对推断程序 $\mathcal{I}$ 的应用会产生一个区间 $\mathbf{I}^q$，其中 $\theta^q \in \mathbf{I}^q$，使得 $\mathcal{O}_q$ 可以被理解为从一个分布 $\bar{P}_{\theta^q} \in \mathcal{P}$ [2]中抽取的一组样本。相关的观测值集合由 $\mathcal{O}_q = f_{\mathcal{S}}(s_q)$ 给出，其中$f_{\mathcal{S}}$： $\mathcal{S} \to 2^{\mathcal{O}}$ 刻画了一个选择 程序。

例如，考虑不同经济体 i 的情况，其中每一个都由一个包含少数几个

[1]使用一部分观测值作为对照组的另一个原因，与大总体定律这一概念有关。Meng (2018) 指的是数据之间的差异 *quantity* 和数据 *quality*。[2]也称为 *Transitional inference*，如 Li 和 Meng (2021) 所述。

¯图 2：模拟对照组 $\in \mathcal{O}_q$



¯通过 $f_{\mathcal{S}}$检测到的相关集合 $\mathcal{O}_q$被扩展为更大的集合 $\mathcal{O}_q$。基于新对照组的¯推断差异可以表示为 $\Delta_{\theta^q} = E(P_{\theta^q}) - E(P_{\theta^q})$。

宏观经济变量 $o_i = \langle x_i, y_i \rangle$，其中 y_i 是国家 i的国内生产总值。我们可以询问，对于任何给定的经济体 $\bar{o} \notin \mathcal{O}$，仅知道 $\bar{x}$，其五年后的预期国内生产总值是多少。现在假设潜变量是某个经济体中主导产业的生产率。那么，我们可以将$\mathcal{O}$ 中所有生产率与 $\bar{o}$ 相似的国家分组，并基于这一选择生成一个对照组类别。

再举一个例子，我们可以考虑一个包含课程项目、学生和成绩的数据集。对于一名新学生，我们可以提出不同的问题。例如，一名新学生需要多长时间才能完成学业？她更换专业的概率是多少？或者她辍学的概率是多少？除此之外，对于这名特定学生，还有许多其他可能的查询，都可以通过数据集中的观测值得到回答。当然，不同的查询可能需要不同的对照组。这一点对于变量和观测值都可能成立。因此，s_q 潜变量将给出在给定 q的情况下相关的观测值部分，并让我们估计 $f_{\mathcal{S}}$，以便模拟新的对照组并提高推断的稳健性。

要以这种方式继续，我们需要针对每个 q 指定与查询相关的潜变量（或一组潜变量）。然后，我们必须利用现有的 $\mathcal{O}$ 上的某种代理指标或测量来区分其取值范围。这样便能得到相关观测的类别。为简单起见，我们可以假设查询只有两种类型，即 q_A 和 q_B，而潜变量 s_q 定义了 $\mathcal{O}_{q_A}$ 和 $\mathcal{O}_{q_B}$。之后可以将这一设置推广到查询类型多于两种的情况。以学生为例，我们假设有两类学生，并用一个潜变量来区分（例如，该变量可以反映社会经济优势或认知优势）。这样，每个 q 都会对应相应类型的学生，并根据 $f_{\mathcal{S}}$映射到该类型学生的观测值类别。

如前所述，类别 $\mathcal{O}_q$ 是 $\mathcal{O}$ 中与回答 相关 的条目 q。但是，获得这些相关观测后，进行 稳健 推断需要生成 不在 $\mathcal{O}_q$中的新对照组。为此，需要创建 $\langle x, y\rangle$ 与 $\mathcal{O}_q$ 中的条目相似的配对，但不能假设它们与后者共享相同的潜变量取值。这意味着只能使用条目的 可观测 特征$o \in \mathcal{O}_q$ 来创建 $\mathcal{O}_q$中的虚构对照组，例如 Delbianco 等人（2021 年）的第 5 节中的示例。

3 框架

如前所述，我们的程序借鉴了不同方法中的要素：

- Lei 等人(2018) 提出了三种保形预测方法的版本。其中一种基于Vovk 等人(2022) 的原始公式。其他两种是分割保形预测和刀切法替代方案。

- Lei 等人（2018）还提供了一个**R语言**代码，用于进行无分布假设的预测，我们对其进行了修改，以便应用于我们的项目。

- 我们同时使用经典线性方法和一种基于弹性网络的方法。或者，我们也可以纳入高斯核回归，以实现非参数方法。还可以利用 **R语言** 软件包 `Caret` (Kuhn(2008))，该方法使用所生成的估计值和预测值，可以方便地应用其他方法，原因还包括以下几点：1. 它会执行前期训练阶段，通过交叉验证结果来选择有意义的特征。它甚至允许应用不同于保形预测的方式来拆分训练数据库中的条目。这使得我们能够更精细地选择 n_r 相关条目和 $2n_r$ 对照组。2.它允许从近五十种预测方法中进行选择，其中包括 GLM、核回归、分位数回归、随机森林等。但是，如前所述，在第一次运行中我们仅使用回归和 LASSO。

- 我们练习中使用的参数为： α 用于置信区间， ρ 用于分割保形预测中的拆分， γ 用于余弦相似度（对于百分位数相似度，我们还使用 α）。

- 我们增加了两个预测阶段，使用相关性的概念来进行原始展示的个体推断（Delbianco等人 （2021）：

1. 第一阶段根据每个新观测值与数据库其余部分的相似程度，评估数据库中各条目的相关性。当p 较小时，相关性根据观测值尾部与数据库中观测值尾部之间的距离百分比来定义。否则，当尾部的维度较大时（尤其是当 $p > n$ 时），我们采用基于余弦的度量。也就是说，相应尾部之间的余弦越接近 1，某条目对于新观测值的相关性就越高。可以基于相关条目进行预测：使用标准预测方法，在原始数据库的相关条目子集上预测新观测值头部的一个可能值。

2. 第二类预测基于使用上一阶段收集的相关条目生成对照组。对于查询 X_0中的每个单独条目，我们通过向其特征添加少量噪声来生成另一个尾部-头部向量。因此，给定 n_r 个相关条目（$n_r << n$），我们在此阶段获得 $2n_r$ 个对照向量。

现在我们将给出算法的伪代码，该算法可为查询点生成预测区间。请注意，为了简洁起见，我们不会明确说明从 $\mathcal{M}$ (Lei 等人 (2018) 的方法中借鉴的步骤和参数，这些方法已在**R语言**程序包保形推断[3]) 中实现， $\mathcal{A}$ (在`Caret` 程序包[4]中实现)以及 $\mathcal{S}$中的相似度函数。

算法 1 个性化保形预测

输入： X, Y, X_0, 回归方法 $\mathcal{A}$, 相似度方法 $\mathcal{S}$, 保形方法 $\mathcal{M}$, 失覆盖率 α, 相似度强度 γ**输出**：$X_0\alpha \leftarrow \alpha_i \in \{0;\ 1\}\gamma \leftarrow \gamma_i \in \{0;\ 1\}\mathcal{A} \leftarrow \mathcal{A}_i$ 中每个元素的预测区间

$\mathcal{M} \leftarrow \mathcal{M}_i$ ▷ 保形、拆分、Jackknife
$\mathcal{S} \leftarrow \mathcal{S}_i$ ▷ OLS、LASSO、核百分位数、余弦
对于每个 $x \in X_0$ **执行 1.1** 使用 y_0 和 $\mathcal{A}$ 预测 X**1.2** 使用残差和 $\mathcal{M} \rightarrow$ 区间（$\mathcal{C}$）**2.1** 使用 γ 和 $\mathcal{S} \rightarrow X_{rel}$**2.2** 使用 y_0 和 $\mathcal{A}$ 预测 X_{rel}**2.3** 使用残差和 $\mathcal{M} \rightarrow$ 相关区间（$\mathcal{C}_{rel}$）**3.1** 模拟 X_{sim}，模仿 X_{rel} 的分布，并使用 $nsim = nrel$**3.2** 使用 y_0 和 $\mathcal{A}$ 预测 X_{sim}**3.3** 使用残差和 $\mathcal{M} \rightarrow$ 相关模拟区间（$\mathcal{C}_{sim}$）**每个循环结束** $x \in X_0$**返回：** $\mathcal{C}$, $\mathcal{C}_{rel}$, $\mathcal{C}_{sim}$

该算法的示意图如图3所示。我们可以看到，存在

[3] `https://github.com/ryantibs/conformal`
[4] `https://topepo.github.io/caret/index.html`

三条路径，分别指向 $\mathcal{C}$、 $\mathcal{C}_{rel}$ 或 $\mathcal{C}_{sim}$。每条路径都可以独立执行，尽管后两条路径共享一个相关观测的检测阶段。

在图 3 中，需要作出的主要决策由红色圆圈表示。所作选择中的每一项变更都可能导致不同的结果。原则上，我们不知道结果对所作选择的细微差异有多稳健，但其中一些影响很容易预测。例如，较低的显著性水平 α会产生更长的区间，而较高的相似度水平 γ 会减少相关观测的数量。除此之外，估计方法、保形推断类型和相关性准则的实际选择取决于具体问题以及用户对这些问题的评估。

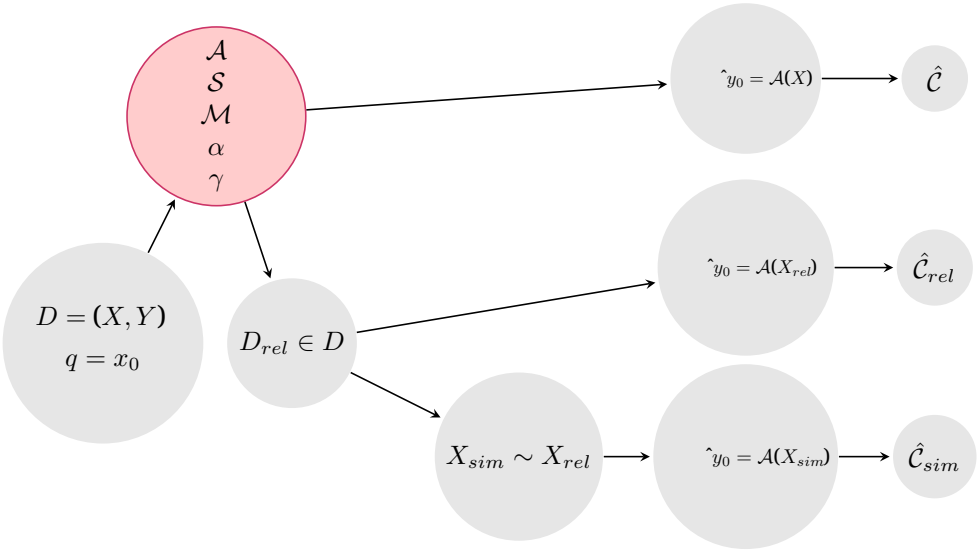


图3：算法 1 的示意图。

红色圆圈框出了算法的初始化阶段，在该阶段会选择估计方法、保形推断的类型以及相关性准则。这些决策会影响所有输出。从这一点开始，可以采取三条备选路径。第一条路径不涉及相关数据，并产生保形预测的标准结果，即 $\mathcal{C}$。第二条和第三条路径要求从 D 中选择一部分与查询 x_0相似的观测值。此后，两种可能的选择分别是：仅使用该子集来获得置信区间 $\mathcal{C}_{rel}$ ，或者生成一个合成样本，将相关数据的类别扩大，以获得 $\mathcal{C}_{sim}$。在算法 1中，这些备选路径分别用 **1**、**2** 和 **3**表示，其各自的步骤为（**1.1**、**1.2**、**2.1**等）。

为了比较所选不同方法的结果，我们采用四个指标。其中两个指标， A 和B，评估的不是区间，而是目标变量的预测值。

A. **与预测值的距离**：真实值 y_0 与 $\mathcal{A}$ 在 X、 X_{rel} 或 X_{sim}上预测的值之间差值的绝对值。显然，A 越小，预测就越准确。

B. **分布 偏离预测值的程度，以 y_0 的百分比表示**：其定义为 $\frac{A}{y_0}$ 表示偏离程度 — is

预测值与实际值之间的差异与后者成正比。在这种情况下，不一定是越小越好。

C. **区间长度**：算法 1 生成的区间上限与下限之差。区间长度取决于 α，但在相同的显著性水平下，区间越短，表示对预测值的不确定性越低。

D. **与预测值的标准化距离**：定义为 $\frac{A}{C}$。它表示误差相对于区间长度的大小。

另外两个指标也可能有助于了解算法 1 生成的区间质量，分别是 覆盖率和 超额 (见 Seedat 等人 (2023))。在第 4 节报告的练习中，所有方法在置信区间内外的预测结果都相似，因此这些指标可能无法提供额外信息。但对于规模更大的异质数据库，这些指标可能会提供更多信息。

4 实证设置

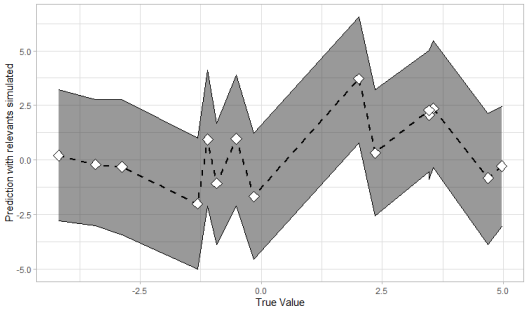
在初步探索中，我们生成了一个包含 750 个观测值的数据集，其中每三分之一的观测值对应一个不同且异质的数据生成过程。相应的设置见表 1。

我们还增加了数据库条目尾部的变量数量（p）。模型的描述见表 2。

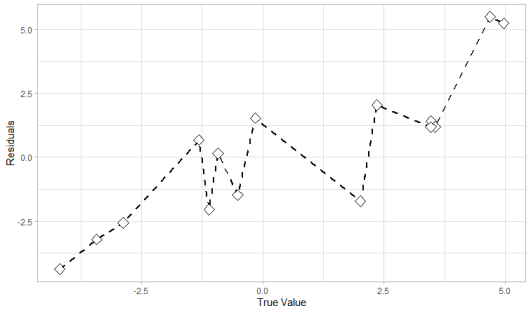
表 3 和 4 展示了在表 1的模拟数据上运行算法 1 的结果。这些结果以及使用高斯核和 LASSO 得到的结果的指标列于表5、6 和 7中。

图 4、5 和 6 表明，预测区间与真实值相适应。值得注意的是，真实值分布尾部的残差会增大。这并不十分令人意外，因为我们没有使用 y_0 来推断用于生成最终预测值的新数据的区间。

图 7 使用较小的数据集绘制，用于比较三种保形预测方法。可以看出，对于用作查询点的三个值，这些方法之间差异不大。

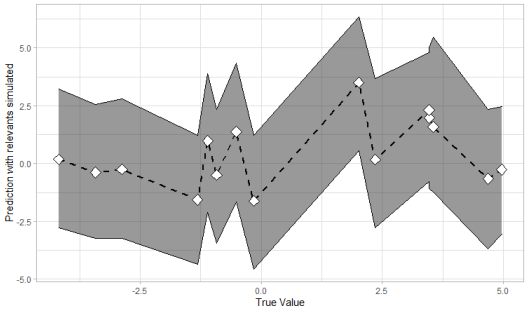


(a) 预测

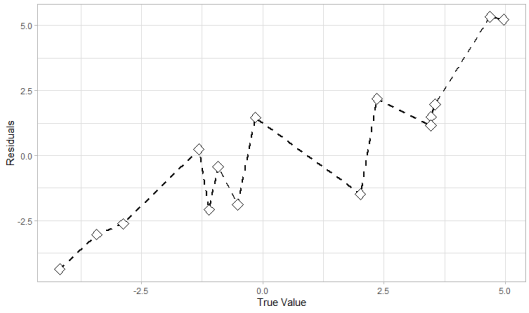


(b) 残差

图 4：使用模拟相关对照组和 OLS 的保形预测

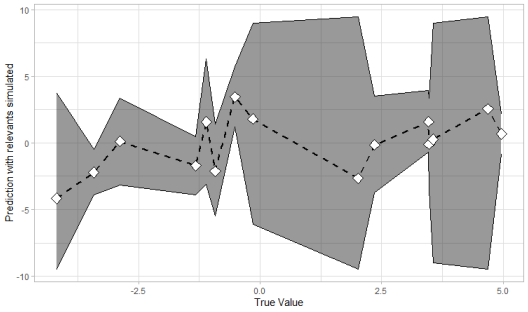


(a) 预测

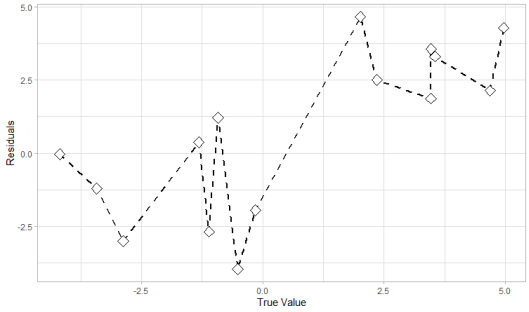


(b) 残差

图 5：使用模拟的相关对照组和 LASSO 进行保形预测

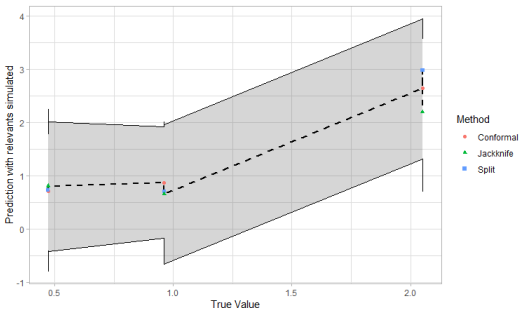


（a）预测值

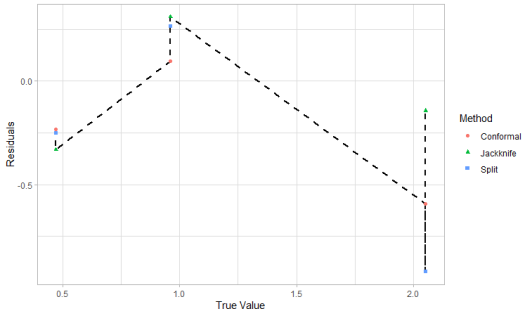


（b）残差

图 6：使用模拟相关对照组和核回归进行保形预测



（a）预测值



（b）残差

图 7：使用模拟相关对照组和 LASSO 的三种方法结果

5 讨论

如前所述，预测区间会针对每个查询进行调整。点预测更多地取决于所使用的数据，而不是所选择的保形预测方法。点值预测往往非常保守，因此残差在左侧趋于负值，在右侧趋于正值。在大多数情况下，预测区间能够涵盖真实值。

当然，这里通过实证分析的案例都非常简单。它们不允许我们穷尽地探究算法 1初始化阶段各项选择所带来的全部后果。但我们可以先作出这样的观察：在保形预测方法中，无法确定一个明显的优胜者，因为这三个变体的平均结果相近，尽管分割保形预测的计算成本较低。

我们的方法仅使用相关数据和模拟数据来生成自适应区间，从而确保预测值的稳- 健性。这种分而治之策略特别适用于大型异质数据集。数据子集使变量的预测值具有个性化，而这些变量的具体信息并不可用。

作为额外的好处，我们可以针对每个查询获得相关数据的分布。这捕捉了一个潜变量（有助于预测与查询对应的区间），并提供有关同一“类别”中个体的有用信息。

不利的一面是，我们看到残差在实际值分布的尾部更为分散。这是应用保形方法所带来的结果；而在Seedat 等人（2023）的自监督方法中，残差则用于训练系统，以改进其预测。

还需要进一步研究，以更精确地评估该方法的优缺点，并将其与文献中介绍的其他程序联系起来，从而加以完善。可能的研究方向包括：

- 使用更大的 p 或更高维的设置进行更多实验。
- 将个体结果从相关群体（甚至从非相关群体）中边缘化出来，如Zhou 等人(2023)。
- 探讨我们的个性化策略是否与 分而治之版本兼容，该版本由 Chen 等人 (2021) 讨论。

- 利用残差在区间两端生成修正值，以降低 认知不确定性（Alaa 等人提出的边界处方法 Alaa 等人 (2023）。

参考文献

Alaa, A. M., Hussain, Z. 和 Sontag, D.（2023）。保形化无条件分位数回归。*arXiv* 预印本 *arXiv:2304.01426*。

Alaa, A.M.和 VanDerSchaar,M.（2017）。使用多任务高斯过程进行个体化治疗效果的贝叶斯推断。神经信息处理系统进展，30。

Barber，R. F.，Candes，E.J.，Ramdas，A. 和 Tibshirani，R. J.（2021）。使用刀切法进行预测推断 +。统计年刊，49(1)：486–507。

Cai, C., Chen,R.,and Xie, M.-g. (2021).个体化群组学习。《美国统计学会会刊》，第1–17页。

Chen,X., Cheng,J. Q.,and Xie, M.-g. (2021).分而治之方法用于大数据分析. *arXiv* 预印本 *arXiv:2102.10771*.

Chernozhukov,V.、Chetverikov,D.、Demirer,M.、Duflo, E.、Hansen, C. 和 Newey,W.（2017）。治疗效应的双重/去偏/Neyman 机器学习。美国经济评论，107(5)：261– 65。

Delbianco, F.、Fioriti, A.和 Tohm´e,F.（2021）。回答个体查询的方法：寻找相关且稳健的对照组。行为计量学，48(2)：259–282。

Infanger, D. 和 Schmidt-Trucks¨ass, A.（2019）。P 值函数：一种被低估的研究结果呈现方法，有助于促进定量推理。医学统计学，38(21):4189–4197。

Kuhn, M.（2008）。使用 caret 软件包在R语言 中构建预测模型。统计软件杂志，28:1–26。

Lei、J.、G'Sell、M.、Rinaldo、A.、Tibshirani、R. J.和Wasserman、L.（2018）。用于回归的无分布预测推断。《美国统计学会会刊》，113(523)：1094–1111。

Li, X. 和 Meng, X.-L.（2021）。一种用于逼近无限 p、零 n 的多分辨率理论：过渡性推断、个体化预测，以及一个没有偏差-方差权衡的世界。《美国统计学会会刊》，116(533):353–367。

Meng, X.-L.（2018）。大数据中的统计学天堂与悖论（I）：大规模总体定律、大数据悖论与2016年美国总统选举。应用统计年鉴，12(2):685–726。

Schweder, T. 和 Hjort, N. L.（2016）。置信度、似然、概率，第41卷。剑桥大学出版社。

Seedat, N., Jeffares, A., Imrie, F. 和 van der Schaar, M.（2023）。利用自监督学习改进自适应保形预测。*arXiv* 预印本 *arXiv:2302.12238*。

Shen, J., Liu, R. Y. 和 Xie, M.-g.（2018）。带置信度的预测——预测推断的通用框架。《统计规划与推断杂志》，195：126–140。

Shen, J., Liu, R. Y. 和 Xie, M.-g.（2020）。i fusion：个体化融合学习。《美国统计学会会刊》，115（531）：1251–1267。

Tran, T., Pham, T., Carneiro, G., Palmer, L. 和 Reid, I.（2017）。一种用于学习深度模型的贝叶斯数据增强方法。神经信息处理系统进展，30。

Vovk, V.、Gammerman, A. 和 Shafer, G.（2022）。随机世界中的算法学习。施普林格。

Xie, M.-g. 和 Singh, K.（2013）。置信分布：参数的频率学派分布估计量：综述。《国际统计评论》，81(1)：3–39。

Xie, M.-g. 和 Wang, P.（2022）。用于有限样本与大样本推断的 Repro 样本方法。*arXiv* 预印本 *arXiv:2206.06421*。

Xie, M.-g. 和 Zheng, Z.（2022）。保形预测和预测分布函数中的稳态现象。《国际近似推理杂志》，141：131–145。

Zhou, Z., Li, D., Huh, D., Xie, M., 和 Mun, E.-Y.（2023）。过多零值计数型结果统计模型性能的模拟研究：聚焦于边际化模型。*arXiv*预印本 *arXiv:2301.12674*。

附录

设置 A 单变量	设置 B 两个变量	设置 C 异方差
$n_1 = 250$	$n_2 = 250$	$n_3 = 250$
u_1 =rnorm(n_1) x_{1a} = rnorm(n_1,1,1) x_{2a} = rnorm(n_1,2,1)（无关） $y_a = 0.5 * x_{1a} + u_1$	u_2 =rnorm(n_2) x_{1b} = rnorm(n_2,3,1) x_{2b} = rnorm(n_2,2,2) $y_b = 0.5 * x_{1b} + 0.33 * x_{2b} + u_2$	x_{1c} = rnorm(n_3,1,1) u_3 = rnorm(n_3) $* \frac{x_{1c}}{2}$ x_{2c} = rnorm(n_3,3,2) $y_c = 0.5 * x_{1c} * x_{1c} + 0.33 * x_{2c} + u_3$
用于 A的新数据 x_{01a} = rnorm(1,1,1) x_{02a} = rnorm(1,2,2) $y_{0a} = 0.5 * x_{01a}$ + rnorm(1)	新数据用于 B x_{01b} = rnorm(1,3,1) x_{02b} = rnorm(1,2,2) $y_{0b} = 0.5 * x_{01b} + 0.33 * x_{02b}$ + rnorm(1)	新数据用于 C x_{01c} = rnorm(1,1,1) x_{02c} = rnorm(1,3,2) $y_{0c} = 0.5 * x_{01c} + 0.33 * x_{02c}$ + rnorm(1) $* \frac{x_{01c}}{2}$

表 1：大型 n 和小型 p 模拟数据集（使用 R语言中的函数 e)

DGP_1:	DGP_2:	DGP_3:
12个变量 $N(0,1)$ 和 2个系数 $=0$	特征均值的偏移	系数均值的偏移
$n_1 = 100; p_1 = 12;\ s_1 = 2$	$n_2 = 100; p_2 = 12;\ s_2 = 2$	$n_3 = 100; p_3 = 12;\ s_3 = 2$
$x_1 =$ matrix(rnorm($n_1 * p_1$) n_1, p_1)	$x_2 =$ matrix(rnorm($n_2 * p_2$, 均值 $= 1$) n_2, p_2)	$x_3 =$ matrix(rnorm($n_3 * p_3$) n_3, p_3)
$\beta_1 =$ c(rnorm(s_1), rep(0, $p_1 - s_1$))	$\beta_2 =$ c(rnorm(s_2), rep(0, $p_2 - s_2$))	$\beta_3 =$ c(rnorm(s_3, 均值 $= 1$), rep(0, $p_3 - s_3$))
$y_1 = x_1 * \beta_1 +$rnorm(n_1)	$y_2 = x_2 * \beta_2 +$rnorm(n_2)	$y_3 = x_3 * \beta_3 +$rnorm(n_3)
$n_{01} = 5$	$n_{02} = 5$	$n_{03} = 5$
$x_{01} =$ matrix(rnorm($n_{01} * p_1$) n_{01}, p_1)	$x_{02} =$ 矩阵(rnorm($n_{02} * p_2$, 均值 $= 1$) n_{02}, p_2)	$x_{03} =$ matrix(rnorm($n_{03} * p_3$) n_{03}, p_3)
$y_{01} = x_{01} * \beta_1 +$rnorm($n_{01}$)	$y_{02} = x_{02} * \beta_2 +$rnorm($n_{02}$)	$y_{03} = x_{03} * \beta_3 +$rnorm($n_{03}$)
$D_1 =$ as.data.frame(cbind(y_1, x_1))	$D_2 =$ as.data.frame(cbind(y_2, x_2))	$D_3 =$ as.data.frame(cbind(y_3, x_3))
$x_1 =$as.data.frame(x_1)	$x_2 =$as.data.frame(x_2)	$x_3 =$as.data.frame(x_3)

表 2：较小 n 和较大的 p 模拟数据集（使用 R 语言中的函数）

余弦	保形的			拆分			刀切法		
变量	1	2	3	1	2	3	1	2	3
y0	0.96	2.05	0.47	0.96	2.05	0.47	0.96	2.05	0.47
pred	0.96	2.59	0.95	0.96	2.79	0.9	2.06	0.71	0.98
predr	0.88	2.43	0.84	0.79	2.81	0.75	0.67	2.19	0.84
predrs	0.88	2.41	0.84	0.91	2.63	0.9	0.69	2.2	0.8
predl	0.96	2.59	0.95	0.96	2.59	0.95	2.05	0.71	0.99
predlr	0.88	2.43	0.84	0.79	2.81	0.75	0.64	2.08	0.78
predlrs	0.9	2.47	0.86	0.91	2.63	0.9	0.65	2.19	0.8
lo	−0.36	1.36	−0.36	−0.83	1.01	−0.89	0.22	−1.13	−0.85
lor	0.65	0.79	1.08	−1.28	0.36	−0.88	−0.73	0.59	−0.48
lors	−0.36	1.08	−0.36	−0.59	0.49	−0.08	−0.63	0.71	−0.42
lol	−0.36	1.36	−0.36	−0.81	0.83	−0.81	0.22	−1.12	−0.84
lolr	0.5	0.79	1.08	−1.28	0.37	−0.88	−0.73	−0.41	−0.62
lolrs	−0.36	1.08	−0.36	−0.59	0.49	−0.08	−0.67	0.71	−0.42
up	2.22	3.8	2.22	2.74	4.58	2.69	3.89	2.54	2.82
upr	3.08	3.51	3.51	2.96	4.52	2.39	2.08	3.79	2.15
uprs	2.08	3.8	2.08	2.39	4.51	1.88	2	3.69	2.03
upl	2.22	3.8	2.22	2.72	4.36	2.71	3.88	2.54	2.82
uplr	3.08	3.51	3.51	2.96	4.53	2.39	2	4.57	2.19
uplrs	2.22	3.8	2.08	2.39	4.51	1.88	1.95	3.57	2.01

表3：余弦相关性

注：$pred$、lo 和up 分别表示预测值、区间下限和上限。相应地，r 表示仅使用相关观测值的情况，而 rs 表示同时纳入相关观测值和模拟观测值的情况。最后，l 和k 分别表示使用 LASSO 回归和核估计的情况。

百分位数	保形的			拆分			刀切法		
变量	1	2	3	1	2	3	1	2	3
y0	0.96	2.05	0.47	0.96	2.05	0.47	0.96	2.05	0.47
pred	0.96	2.59	0.95	0.96	2.79	0.9	2.06	0.71	0.98
predr	0.84	2.57	0.69	0.43	2.05	0.58	0.67	2.19	0.84
predrs	0.86	2.56	0.68	0.7	2.97	0.72	0.69	2.2	0.8
predl	0.96	2.59	0.95	0.96	2.59	0.95	2.05	0.71	0.99
predlr	0.84	2.57	0.69	0.43	2.02	0.58	0.64	2.08	0.78
predlrs	0.87	2.64	0.7	0.7	2.97	0.72	0.65	2.19	0.8
lo	−0.36	1.36	−0.36	−0.34	1.5	−0.39	0.22	−1.13	−0.85
lor	0.76	0.97	−0.28	−0.21	1.25	−0.59	−0.73	0.59	−0.48
lors	−0.25	1.2	−0.41	−0.34	0.88	−0.79	−0.63	0.71	−0.42
lol	−0.36	1.36	−0.36	−0.31	1.33	−0.32	0.22	−1.12	−0.84
lolr	1.05	1.2	−0.47	−0.26	1.25	−0.6	−0.73	−0.41	−0.62
lolrs	−0.18	1.31	−0.41	−0.34	0.88	−0.79	−0.67	0.71	−0.42
up	2.22	3.8	2.22	2.25	4.09	2.19	3.89	2.54	2.82
upr	2.92	3.49	1.91	2.1	4.09	1.75	2.08	3.79	2.15
uprs	1.91	3.94	1.79	2	3.9	2.24	2	3.69	2.03
upl	2.22	3.8	2.22	2.22	3.86	2.21	3.88	2.54	2.82
uplr	3.29	3.83	1.72	2.17	4.09	1.76	2	4.57	2.19
uplrs	1.91	3.94	1.79	2	3.9	2.24	1.95	3.57	2.01

表 4：百分位相关性

注释：此处我们使用与表格中相同的缩写3

短数据

百分位数	通用	保形	拆分	刀切法
差分预测	0.57	0.34	0.39	0.98
差分预测 r	0.25	0.28	0.21	0.26
差分预测 rs	0.33	0.27	0.48	0.25
差分预测 l	0.55	0.34	0.34	0.98
差分预测 lr	0.24	0.28	0.22	0.22
差分预测 lrs	0.35	0.31	0.48	0.26
差分预测 k	0.32	0.3	0.31	0.34
diffpredkr	0.24	0.2	0.24	0.28
diffpredkrs	0.25	0.19	0.27	0.28
%pred	60.25	42.49	42.38	95.87
%predr	30.52	27.57	26.09	37.92
%predrs	34.42	26.03	41.81	35.41
%predl	60.31	42.49	42.56	95.89
%predlr	29.28	27.57	26.65	33.62
%predlrs	35.9	29.26	41.91	36.54
%predk	38.37	36.05	36.49	42.58
%predkr	26.22	23.44	27.83	27.39
%predkrs	27.79	23.29	32.62	27.45
int	2.93	2.53	2.59	3.67
intr	2.56	2.29	2.49	2.89
intrs	2.62	2.37	2.8	2.69
intl	2.91	2.53	2.53	3.66
intlr	2.8	2.35	2.54	3.5
intlrs	2.58	2.31	2.8	2.64
intk	2.56	2.53	2.49	2.65
intkr	2.95	2.56	2.61	3.66
intkrs	2.45	2.29	2.65	2.41
ab	0.6	0.63	0.65	0.52
abr	0.61	0.71	0.59	0.53
abrs	0.55	0.58	0.55	0.53
abl	0.6	0.63	0.63	0.52
ablr	0.62	0.76	0.59	0.5
ablrs	0.55	0.59	0.55	0.51
abk	0.62	0.61	0.62	0.63
abkr	0.57	0.63	0.54	0.55
abkrs	0.54	0.54	0.51	0.58

表5：短数据百分位数

注意：差分预测 表示预测值与真实值之间的差值。$\%pred$ 表示相对于真实值的同一差值。int 是区间长度（上限减去下限）。最后，ab 表示误差相对于区间长度的大小。

短数据

余弦	通用	保形	拆分	刀切法
差分预测	0.57	0.34	0.39	0.98
差分预测 r	0.31	0.28	0.4	0.26
差分预测 rs	0.29	0.27	0.35	0.25
差分预测 l	0.55	0.34	0.34	0.98
差分预测 lr	0.3	0.28	0.4	0.22
差分预测 lrs	0.3	0.29	0.35	0.26
差分预测 k	0.32	0.3	0.34	0.31
diffpredkr	0.6	0.8	0.57	0.44
diffpredkrs	0.93	1.1	0.58	1.11
%pred	60.25	42.49	42.38	95.87
%predr	37.08	35.18	38.15	37.92
%predrs	37.27	34.79	41.6	35.41
%predl	60.31	42.49	42.56	95.89
%predlr	35.65	35.18	38.15	33.62
%predlrs	38.18	36.4	41.6	36.54
%predk	38.41	36.05	42.7	36.49
%predkr	89.13	102.14	99.06	66.2
%predkrs	144.67	153.04	99.05	181.92
int	3.26	2.53	3.58	3.67
intr	3.1	2.53	3.89	2.89
intrs	2.74	2.53	2.99	2.69
intl	3.24	2.53	3.52	3.66
intlr	3.33	2.58	3.89	3.5
intlrs	2.74	2.58	2.99	2.64
intk	2.56	2.53	2.64	2.49
intkr	2.93	2.86	3.41	2.52
intkrs	2.91	3.37	2.68	2.67
ab	0.59	0.63	0.61	0.52
abr	0.65	0.88	0.55	0.53
abrs	0.57	0.59	0.6	0.53
abl	0.58	0.63	0.6	0.52
ablr	0.64	0.87	0.55	0.5
ablrs	0.57	0.6	0.6	0.51
abk	0.62	0.61	0.63	0.62
abkr	0.62	0.37	0.54	0.94
abkrs	0.73	0.62	0.58	0.99

表6：短数据余弦

注释：这里我们使用与表格中相同的缩写 5

长数据

余弦	通用	保形	拆分	刀切法
差分预测	2.8	2.47	2.58	3.34
差分预测 r	2.75	2.27	2.64	3.34
差分预测 rs	2.8	2.29	2.76	3.34
差分预测 l	2.81	2.48	2.6	3.35
差分预测 lr	2.73	2.29	2.61	3.28
差分预测 lrs	2.78	2.34	2.69	3.33
差分预测 k	2.05	1.5	2.32	2.33
diffpredkr	1.91	1.7	1.87	2.17
diffpredkrs	2.1	1.77	2.39	2.14
%pred	179.03	165.13	105.96	266.02
%predr	168.86	157.93	115.92	232.74
%predrs	183.81	157.16	165.13	229.12
%predl	176.34	165.17	98.11	265.75
%predlr	173.31	135.75	109.4	274.78
%predlrs	174.9	160.28	136.48	227.92
%predk	161.98	132.38	183.64	169.93
%predkr	131.49	166.7	109.41	118.35
%predkrs	154.95	126.67	215.02	123.15
int	5.13	5.86	5.96	3.57
intr	5.03	5.51	5.95	3.64
intrs	4.31	5.87	3.52	3.54
intl	5.14	5.69	6.38	3.36
intlr	6.78	5.76	6.24	8.33
intlrs	4.31	5.94	3.52	3.48
intk	5.43	4.12	6.71	5.45
intkr	6.05	5.27	5.58	7.31
intkrs	5.97	1.77	9.41	6.73
ab	0.6	0.42	0.43	0.94
abr	0.59	0.41	0.45	0.92
abrs	0.71	0.39	0.8	0.94
abl	0.61	0.44	0.41	1
ablr	0.41	0.4	0.42	0.4
ablrs	0.75	0.39	0.78	1.09
abk	0.38	0.36	0.35	0.43
abkr	0.33	0.33	0.35	0.3
abkrs	0.63	1.21	0.36	0.32

表 7：长数据余弦

注释：这里使用与表5