arXiv:2301.02692v1 [stat.ME] 6 Jan 2023

低信噪比下的保序再校准

Mario V. W¨uthrich* Johanna Ziegel†

2023年1月10日版本

摘要

保险定价系统应满足自动校准性质，以确保不同价格群组之间不存在系统性交叉融资。回归模型往往未经自动校准。我们提出对给定的回归模型应用保序再校准，以确保其自动校准。我们的主要结果证明，在低信噪比下，这一保序再校准步骤能够带来可解释的定价系统，因为由此得到的保序再校准回归函数复杂度较低。

关键词。 自动校准、保序回归、保序再校准、低信噪比、交叉融资、算法解决方案、深度神经网络、可解释性。

1 引言

本文将探讨保险定价中两个看似无关的问题。首先，保险定价系统不应在不同价格群组之间存在系统性的交叉融资。系统性交叉融资意味着投资组合的某些部分定价偏低，而这由投资组合其他定价偏高的部分予以弥补。确保定价系统具有自动校准性质，就能避免价格群组之间的系统性交叉融资。我们提议采用保序再校准，将任意回归函数转化为自动校准的定价系统。

我们要解决的第二个问题，是复杂算法模型在保险定价中的可解释性。首先，可以使用任意复杂的回归模型来设计保险定价系统，例如深度神经网络。这类复杂回归模型通常缺乏可解释性，更像是黑箱。因此，人们开发了多种工具来解释这类复杂解决方案，例如 Lundberg– Lee 的 SHAP [22]。由于算法解决方案通常不具备前述自动校准性质，我们提议对算法解决方案进行保序再校准。如果数据中的信噪比较低，那么保序再校准步骤会对协变量空间进行粗略划分，因此，最初采用的算法模型便会得到一个可解释的版本。因此，在低信噪比问题中应用保序再校准，可解释性是一个不错的附带成果；而保险定价情境通常正是如此。

还有其他方法可以通过重新校准步骤实现自动校准；我们提及 Lindholm 等人 [21] 和 Denuit 等人 [8]。这些方法通常需要调整

*RiskLab，苏黎世联邦理工学院数学系，mario.wuethrich@math.ethz.ch
†伯尔尼大学数理统计与精算科学研究所，johanna.ziegel@stat.unibe.ch

超参数，例如使用交叉验证。保序再校准不涉及任何超参数，因为它求解的是一个约束回归问题（确保单调性）。因此，保序再校准具有普适性，因为它也不依赖于布雷格曼损失这一类损失函数中的具体选择。

我们将所提出的方法形式化。下文中，我们假设所有考虑的随机变量都具有有限均值。考虑一个响应变量 Y，它具有协变量信息 $\boldsymbol{X} \in \mathcal{X} \subseteq \mathbb{R}^q$。一般目标是确定（真实的）回归函数 $\boldsymbol{x} \mapsto \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$，该函数描述给定 $\boldsymbol{X}$ 时 Y 的条件均值。通常，这个真实回归函数是未知的，需要根据独立同分布数据 $(y_i,\ \boldsymbol{x}_i)_{i=1}^n$ 来确定，也就是说，需要根据 $(Y,\ \boldsymbol{X})$ 的样本来确定。为此，我们尝试从 $\mathcal{X}$ 上预先选定的函数类中选择一个回归函数 $\boldsymbol{x} \mapsto \mu(\boldsymbol{x})$，使其尽可能准确地逼近条件均值 $\mathbb{E}[Y|\boldsymbol{X} = \cdot]$。通常，无法从数据中捕捉回归函数的所有特征。在金融应用中，经过恰当选择的回归函数 $\mu(\cdot)$ 至少必须满足自动校准性质。

定义1.1 回归函数 μ 对于$(Y, \boldsymbol{X})$ 是自动校准的，当且仅当

$$\mu(\boldsymbol{X}) = \mathbb{E}\left[Y \middle| \mu(\boldsymbol{X})\right], \qquad \mathbb{P}\text{-}a.s.$$

自动校准是精算和金融应用中的一项重要性质，因为它意味着，对于相应的索赔Y，价格群组 $\mu(\boldsymbol{X})$平均而言能够自我融资；也就是说，如果分别用协变量 $\boldsymbol{X} \sim \mathbb{P}$ 和价格群组 $\mu(\boldsymbol{X})$描述投资组合的结构，那么投资组合内部不存在系统性交叉融资。在伯努利情境中，自动校准的早期版本（称为校准良好）由 Schervish [28] 引入统计学界；近年来，Gneiting–Resin [12]对其进行了详细研究。在精算和金融领域，Kr¨uger–Ziegel [17], Denuit 等人 [8], W¨uthrich [30] 以及 Lindholm 等人 [21]强调了自动校准的重要性。

许多回归模型并不满足自动校准性质。然而，有一种简单而强大的方法，我们称之为保序再校准，可以从任意候选函数 $\pi : \mathcal{X} \to \mathbb{R}$ 出发，得到一个（样本内）自动校准的回归函数。我们将保序再校准应用于伪样本 $(y_i, \pi(\boldsymbol{x}_i))_{i=1}^n$，从而得到一个保序回归函数 $\widehat{\mu}$。然后，

$$\widehat{\mu}(\boldsymbol{X}') = \mathbb{E}\left[Y' \middle| \widehat{\mu}(\boldsymbol{X}')\right], \quad \mathbb{P}_n\text{-a.s.}, \tag{1.1}$$

其中 $(Y', \boldsymbol{X}')$ 服从 $(y_i, \boldsymbol{x}_i)_{i=1}^n$ 的经验分布 $\mathbb{P}_n$；详情参见第2.1节。保序回归确定了协变量空间 $\mathcal{X}$ 的自适应划分，而 $\widehat{\mu}$ 则通过对各分区元素上的 y 值取均值来确定。显然，也可以在伪样本 $(y_i, \pi(\boldsymbol{x}_i))_{i=1}^n$ 上使用其他分箱方法来确保 (1.1)，但我们认为保序回归更可取，因为它避免了对调优参数作出主观选择，并且在合理且可验证的假设下能够得到合理的回归函数。保序再校准要发挥作用，唯一的假设是函数 π 能够正确排列条件均值，即当 $\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}_i] \leq \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}_j]$时，我们希望 $\pi(\boldsymbol{x}_i) \leq \pi(\boldsymbol{x}_j)$。

使用保序回归进行重新校准在文献中并非新鲜事。对于二元结果，Zadrozny–Elkan [32], Menon 等人 [23] 此前已提出这一方法，最近

Tasche [29, 第5节3]也提出了这一方法。Balabdaoui 等人 [2] 的单调单指标模型采用了上述相同的策略，但其研究重点与我们的不同。他们具体考虑了候选函数 π的线性回归模型，并将其称为指标。在分布回归中，也就是说，当人们希望确定给定协变量信息 $\boldsymbol{X}$时 Y 的完整条件分布时，Henzi 等人 [13] 建议先估计一个指标函数 π，该函数用于确定条件分布相对于一阶随机占优的排序，然后使用保序分布回归估计条件分布；参见 Henzi 等人 [14]。

作为一项新的贡献，我们表明，保序再校准所产生的划分大小，或许能帮助我们了解再校准回归函数 $\widehat{\mu}$ 的信息含量。此外，保序再校准的划分有助于解释协变量与结果变量之间的联系，尤其是在信噪比较低时——这通常是保险理赔数据的情形。

为了构造候选函数 $\pi : \mathcal{X} \to \mathbb{R}$，可以考虑任意回归模型，例如广义线性模型、回归树、树提升回归模型或深度神经网络回归模型。目标是使 $\pi(\cdot)$ 能够为我们提供条件均值 $\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}_i]$, $i = 1, \ldots\ldots, n$ 的正确排序。详情见第3节。

组织结构。在第2节中，我们正式介绍保序回归，这是一种约束优化问题。通常使用相邻违序池算法（PAV）求解这一约束优化问题，该算法见附录 A.1。我们的主要结果陈述于第2.2节，揭示了保序再校准解的复杂度与数据中的信噪比之间的关系。第3节提供了有关如何使用保序再校准的实践指导；第4节则通过一个常用的保险数据集示例说明我们的结果。在这一节中，我们还介绍了用于解释回归函数的图形工具。最后，我们在第5节总结全文。

2 保序回归

2.1 定义与基本性质

为简便起见，我们假设候选函数 $\pi : \mathcal{X} \to \mathbb{R}$ 不会导致值 $\pi(\boldsymbol{x}_1) \ldots, \pi(\boldsymbol{x}_n)$之间出现并列，并且指标 $i = 1, \ldots, n$ 的选取与排序一致，即 $\pi(\boldsymbol{x}_1) < \ldots < \pi(\boldsymbol{x}_n)$。注2.1说明了如何处理并列情况。对于具有正样本权重 $(w_i)_{i=1}^n$ 的 $\boldsymbol{z} = (y_i, \pi(\boldsymbol{x}_i))_{i=1}^n$，其保序回归是受限最小化问题的解 $\widehat{\boldsymbol{\mu}} \in \mathbb{R}^n$:

$$\widehat{\boldsymbol{\mu}} \;=\; \underset{\boldsymbol{\mu}=(\mu_1,\ldots,\mu_n)^\top}{\arg\min} \sum_{i=1}^{n} w_i \left(y_i - \mu_i\right)^2, \qquad \text{subject to } \mu_1 \leq \ldots \leq \mu_n. \tag{2.1}$$

我们可以将侧约束改写为 $A\boldsymbol{\mu} \geq \mathbf{0}$（逐分量），其中 $A = (a_{i,j})_{i,j} \in \mathbb{R}^{n\times(n-1)}$是元素为 $a_{i,j} = \mathbb{1}_{i=j-1} - \mathbb{1}_{i=j}$ 的矩阵。我们定义 $\boldsymbol{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ 以及（对角）案例权重矩阵 $W = \mathrm{diag}(w_1, \ldots, w_n)$。上述优化问题于是可写为

$$\widehat{\boldsymbol{\mu}} \;=\; \widehat{\boldsymbol{\mu}}(\boldsymbol{z}) \;=\; \underset{\boldsymbol{\mu}:\, A\boldsymbol{\mu}\geq\mathbf{0}}{\arg\min}\; (\boldsymbol{y}-\boldsymbol{\mu})^\top W (\boldsymbol{y}-\boldsymbol{\mu}). \tag{2.2}$$

这表明，保序回归可以通过带有线性侧约束的凸最小化求解。接下来需要验证 (1.1) 中所声称的自动校准性质是否成立。

注2.1 如果值 $\pi(\boldsymbol{x}_1) \ldots, \pi(\boldsymbol{x}_n)$ 中存在并列，例如，对于某个 $i \neq j$, $\pi(\boldsymbol{x}_i) = \pi(\boldsymbol{x}_j)$，我们将 y_i 和 y_j 替换为其加权平均值 $(w_i y_i + w_j y_j)/(w_i + w_j)$，并为其分配权重 $(w_i + w_j)/2$。对于两个以上的并列值，处理过程是类似的。这对应于 Leeuw 等人 [20, 第2节1] 中处理并列情况的第二种方案。

注2.2 Barlow 等人 [3, 定理1.10] 证明，(2.1) 中的平方损失函数可以替换为任意 Bregman 损失函数，$L_\phi(y, \mu) = \phi(y) - \phi(\mu) + \phi'(\mu)(y - \mu)$，而不会改变最优解 $\widehat{\boldsymbol{\mu}}$。这里，$\phi$ 是一个具有次梯度 ϕ'的严格凸函数。Bregman 损失函数是均值的一致损失函数；参见 Savage [27] 和 Gneiting [11,定理 7]。如果 y 和 μ 仅取正值，则与本文相关的 Bregman 损失函数是伽马偏差损失；通过选择$\phi(x) = -\log(x)$，它等价于由此产生的 QLIKE 损失；参见 Patton [25]。

最小化问题 (2.2) 的解可通过极小极大公式显式给出，即

$$\widehat{\mu}_i = \min_{\ell=i,\ldots,n} \max_{k=1,\ldots,\ell} \frac{1}{\sum_{j=k}^{\ell} w_j} \sum_{j=k}^{\ell} w_j y_j.$$

尽管极小极大公式在理论上颇具吸引力且很有用，但 Brunk 等人 [6]提出的相关最小下集（MLS）算法无法高效地计算解。相邻违序池算法（PAV）（由 Ayer 等人 [1], 、Miles [24] 和 Kruskal [18],提出）能够快速计算保序回归，并为我们提供关于解的所需见解。在附录 A.1 中，我们详细介绍 PAV 算法。通过将指标集 $\mathcal{I} = \{1, \ldots, n\}$ 适当地划分为（离散）区间，可得到解

$$\mathcal{I}_k = \mathcal{I}_k(\boldsymbol{z}) = \{i_{k-1} + 1, \ldots, i_k\} \qquad \text{for } k = 1, \ldots, K(\boldsymbol{z}), \tag{2.3}$$

这些区间的切分点 $0 = i_0 < i_1 < \ldots < i_K = n$取决于 $\boldsymbol{z}$，且 $K(\boldsymbol{z}) \in \{1, \ldots, n\}$表示离散区间 $\mathcal{I}_k$的数量。区间数量 $K(\boldsymbol{z})$ 以及用于划分 $\mathcal{I}$的切分点 $i_k = i_k(\boldsymbol{z})$、$k = 1$、.....、$K(\boldsymbol{z})$ 取决于观测值 $\boldsymbol{z}$。随后，我们在每个离散区间 $\mathcal{I}_k$ 上得到保序回归参数估计，例如$i \in \mathcal{I}_k$

$$\widehat{\mu}_i = \widehat{\mu}_{i_k} = \frac{1}{\sum_{j \in \mathcal{I}_k} w_j} \sum_{j \in \mathcal{I}_k} w_j y_j, \tag{2.4}$$

另见 (A.5)。因此，在每个块 $\mathcal{I}_k$ 上，我们都有一个常数估计值 $\widehat{\mu}_{i_k}$；保序性质表明，这些估计值沿块指标 $k = 1, \ldots, K(\boldsymbol{z})$ 严格递增，因为这些块是按极大性选取的。我们将 $K(\boldsymbol{z})$ 称为所得保序回归的**复杂度数** 。

图1给出了 $n = 20$ 以及排序 $\pi(\boldsymbol{x}_i) = i$ 对应于 $i = 1, \ldots, n$ 的一个示例。由此得到的（非参数）保序回归函数 $\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}(\boldsymbol{z})$ 仅在观测值处（$\pi(\boldsymbol{x}_i)$）$_{i=1}^{n}$ 唯一确定，可以用阶梯函数进行插值。在图1中，这得到一个具有 $K(\boldsymbol{z}) - 1 = 9$ 个阶梯的阶梯函数，也就是说，我们有 $K(\boldsymbol{z}) = 10$ 个块，而估计回归函数 $\widehat{\mu}$ 仅取 $K(\boldsymbol{z}) = 10$ 个不同的值。这启发我们将 $K(\boldsymbol{z})$ 称为所得阶梯函数的复杂度数，见图1。

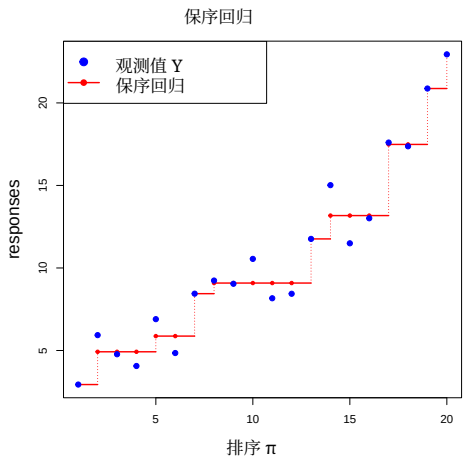


图1：含有 $K(\boldsymbol{z}) = 10$ 个块的保序回归示例。

通过要求单调性，可以自然地将指标 $\mathcal{I}$ 划分为保序块 $\mathcal{I}_k$ 。这不同于 Lindholm 等人 [21] 所考虑的回归树方法。事实上，后者并不要求单调性，而是旨在最小化“普通”的平方损失，例如使用交叉验证来确定最优划分数。在我们的背景下，复杂度数 $K(\boldsymbol{z})$ 完全由单调性要求决定，因此通常会得到不同的结果。

在保险应用中，块 $\mathcal{I}_k \subset \mathcal{I}$ 为我们提供了（经验）价格群组 $\widehat{\mu}_i = \widehat{\mu}_{i_k}$，对于 $i \in \mathcal{I}_k$，而 (2.4) 则导出了 Y 的（样本内）自动校准性质

$$\mathbb{E}\left[Y' \middle| \widehat{\mu}(\boldsymbol{X}') = \widehat{\mu}_{i_k}\right] \;=\; \frac{1}{\sum_{i\in\mathcal{I}_k} w_i} \sum_{i\in\mathcal{I}_k} w_i y_i \;=\; \widehat{\mu}_{i_k}, \tag{2.5}$$

其中，$(Y', \boldsymbol{X}')$服从$(y_i, \boldsymbol{x}_i)$的加权经验分布$_{i=1}^n$，权重为$(w_i)_{i=1}^n$。此外，对整个投资组合求和可得（全局）平衡性质

$$\sum_{i=1}^{n} w_i \widehat{\mu}_i = \sum_{k=1}^{K(\boldsymbol{z})} \sum_{i\in\mathcal{I}_k} w_i \widehat{\mu}_i = \sum_{k=1}^{K(\boldsymbol{z})} \widehat{\mu}_{i_k} \sum_{i\in\mathcal{I}_k} w_i = \sum_{k=1}^{K(\boldsymbol{z})} \sum_{i\in\mathcal{I}_k} w_i y_i = \sum_{i=1}^{n} w_i y_i, \tag{2.6}$$

也就是说，若我们使用协变量 $\boldsymbol{x}_i$，并按照 $w_i\widehat{\mu}_i$为保险单定价，那么平均而言，整体（价格）水平的设定是正确的；此时，权重 $w_i > 0$ 可解释为风险暴露量。

2.2 期望复杂度数的单调性

本节将证明，期望复杂度数 $\mathbb{E}[K(\boldsymbol{z})]$ 是信噪比的递增函数。为此，我们假设响应变量 Y_i服从位置-尺度模型，即假设

$$Y_i = \mu_i + \sigma\epsilon_i, \quad i = 1, \ldots, n, \tag{2.7}$$

噪声项为 ϵ_i，位置参数为 $\mu_i \in \mathbb{R}$，其中 $\mu_1 \leq \ldots \leq \mu_n$，尺度参数为$\sigma > 0$。这里，$\mu_i$ 在上一节中扮演 $\pi(\boldsymbol{x}_i)$ 的角色。参数 $\mu_1, \ldots, \mu_n$ 未知，但已知它们按递增顺序排列。信噪比为

接着由尺度参数 σ 描述，即 σ 较大时信噪比较低，反之亦然。显式的位置-尺度结构 (2.7) 使我们能够分析

$$\boldsymbol{y} \;=\; \boldsymbol{Y}_\sigma(\omega) \;=\; \boldsymbol{\mu} + \sigma\boldsymbol{\epsilon}(\omega) \;=\; (\mu_1,\dots,\mu_n)^\top + \sigma(\epsilon_1,\dots,\epsilon_n)^\top(\omega), \tag{2.8}$$

作为概率空间 $(\Omega,\ \mathcal{F},\ \mathbb{P})$ 中样本点 $\omega \in \Omega$ 的逐点函数，关于 $\sigma > 0$；这类似于 Kingma–Welling [16] 的重参数化技巧，该技巧常用于探索变分自编码器。

在本节中，我们写作 $K(\boldsymbol{y}) = K(\boldsymbol{z})$，因为根据上下文（标记）可以清楚地看出结果变量 $\boldsymbol{y}$ 的排序，因此我们不通过排序函数 $\pi(\cdot)$。

定理 2.3 假设响应变量 $Y_i,\ i = 1,\dots,n$ 遵循位置-尺度模型 (2.7)，其（未知的）位置参数 $\mu_1 < \ldots < \mu_n$ 按序排列，尺度参数为 $\sigma > 0$。那么，对 $\boldsymbol{Y}$ 进行保序回归的期望复杂度数 $\mathbb{E}[K(\boldsymbol{Y})]$ 是 $\sigma > 0$ 的递减函数。如果噪声向量 $\boldsymbol{\epsilon} = (\epsilon_1,\dots,\epsilon_n)^\top$ 的分布在 $\mathbb{R}^n$ 上具有全支撑，那么 $\mathbb{E}[K(\boldsymbol{Y})]$ 关于 σ 严格递减。

定理 2.3 证明，在一种特定但与实际高度相关的模型下，保序回归的复杂度数$K(\boldsymbol{Y})$会随着信噪比降低而平均下降。这隐含地意味着，含噪程度更高、信息比更低的数据会产生粒度较粗的回归函数。因此，如果利用保序回归的划分，通过候选函数 π获得协变量空间 $\mathcal{X}$ 的一个划分，那么无法由 $\pi(\boldsymbol{X})$解释的 $\boldsymbol{Y}$ 噪声越多，该划分的粒度就越粗。进一步讨论也见第3.3节。

据我们所知，我们的结果为保序回归领域的文献作出了新的贡献。虽然我们关注有限样本情形，但也有一项相关研究分析了保序回归函数的复杂度数如何随样本量 n变化，见 Dimitriadis 等人 [9, 引理 3.2]。

在定理 2.3 的表述中，我们假设位置参数严格有序。这一假设简化了证明：我们证明期望复杂度数 $K(\boldsymbol{Y})$关于 σ严格递减。稍作补充记号后，该定理可以推广到允许部分（但并非全部） μ_i 相等的情形。

图 2 展示了一个位置-尺度模型 (2.7) 的示例，其中包含独立同分布的标准高斯噪声，尺度参数分别为 $\sigma = 2$ (左图) 和 $\sigma = 20$ (右图)，两幅图都考虑噪声项 $\boldsymbol{\epsilon}(\omega)$ 中的同一个样本点$\omega \in \Omega$，见 (2.8)。图 2 右侧的复杂度数为 $K(\boldsymbol{y}) = 13$，左侧则为 $K(\boldsymbol{y}) = 46$；选取的样本量为 $n = 100$。

3 用于预测和解释的保序再校准

3.1 预测与估计

为了从独立同分布数据 $\boldsymbol{x} \mapsto \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}](y_i,\ \boldsymbol{x}_i)_{i=1}^n$中确定真实回归函数的自动校准模型，我们提出了一个两步估计程序。首先，我们选择一个回归模型，并使用数据 $y_i,\ \boldsymbol{x}_i{}_{i=1}^n$ 来获得候选函数 $\widehat{\pi}\ \pi$ 的估计值，该候选函数应满足

$$\pi(\boldsymbol{x}) \le \pi(\boldsymbol{x}') \quad\Longleftrightarrow\quad \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}] \le \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}'], \tag{3.1}$$

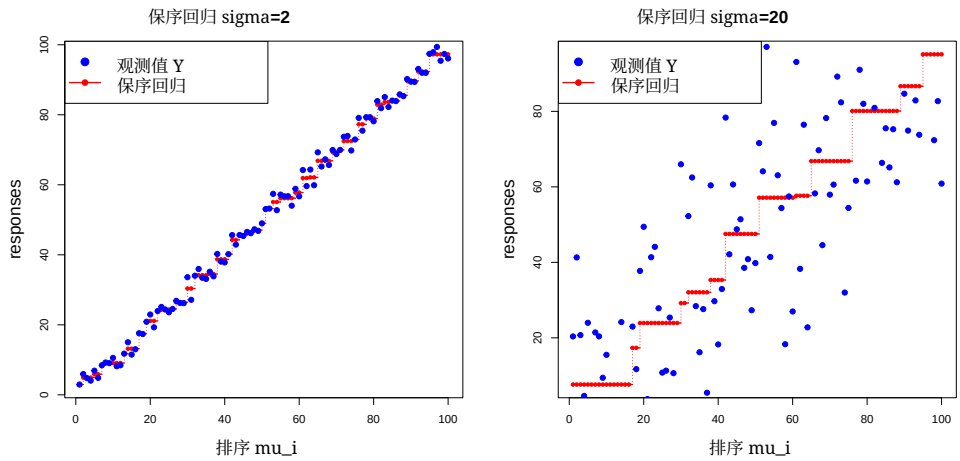


图 2：对位置-尺度类型的保序回归示例：对于相同的样本点 $\omega \in \Omega$，信噪比有所不同：(左图) $\sigma = 2$ with $K(\boldsymbol{y}) = 46$ and (右图) $\sigma = 20$ with$K(\boldsymbol{y}) = 13$。

对于所有 $\boldsymbol{x}$，$\boldsymbol{x}' \in \mathcal{X}$均成立。例如，在第4节的案例研究中，为 π选择了一个深度神经网络模型。为了获得合理的结果，$\hat{\pi}$ 的估计方法不对数据过拟合非常重要。

第二步，我们对伪样本 $(y_i, \hat{\pi}(\boldsymbol{x}_i))_{i=1}^n$ 应用保序回归，从而得到一个定义在 $\{\hat{\pi}(\boldsymbol{x}_i)$ 上的样本内自动校准回归函数 $\hat{\mu}$：$i = 1, \ldots\ldots, n\}$。我们将这第二步称为保序再校准。为了获得新协变量值 $\boldsymbol{x} \in \mathcal{X}$ 的预测，我们计算 $\hat{\pi}(\boldsymbol{x})$，找到满足 $\hat{\pi}(\boldsymbol{x}_i) < \hat{\pi}(\boldsymbol{x}) \leq \hat{\pi}(\boldsymbol{x}_{i+1})$ 的 i，并通过设定 $\hat{\mu}(\boldsymbol{x}) = (\hat{\mu}(\boldsymbol{x}_i) + \hat{\mu}(\boldsymbol{x}_{i+1}))/2$ 进行插值。这种插值可能有利于预测。然而，对于解释和分析，我们更倾向于使用阶梯函数插值，因为这样会得到协变量空间的一个划分，见下文第3.3节和图2。

这种两步估计方法可以被解读为对 Balabdaoui 等人 [2]所研究的单调单指标模型的推广。他们假设真实回归函数具有 $\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}] = \psi(\boldsymbol{\alpha}^\top \boldsymbol{x})$ 的形式，其中 ψ是一个递增函数。与我们提出的方法不同，在他们的方法中，回归模型 π 被固定为线性模型 $\boldsymbol{\alpha}^\top \boldsymbol{x}$。他们针对 $(\psi, \boldsymbol{\alpha})$ 联合进行全局最小二乘估计，但发现计算量很大。作为替代方案，他们提出了一种与我们的方法类似的两步估计程序，但将数据拆分，使得 $\boldsymbol{\alpha}$ 和保序回归分别基于独立样本进行估计。他们发现，如果 $\boldsymbol{\alpha}$ 的估计量具有足够快的收敛速度，那么所得真实回归函数估计量是一致的，且收敛速度为 $n^{1/3}$ 阶。

在分布回归框架中，Henzi 等人 [13] 采用了上述两步估计程序，并以保序分布回归 [14], 替代了第2.1节所述的经典最小二乘保序回归。他们表明，在有数据拆分和无数据拆分两种情况下，只要指标 π 能够以参数收敛速度估计，该程序就能一致地估计给定 $\boldsymbol{X}$时 Y 的条件分布。对于 Henzi 等人 [13] 考察的应用，有无数据拆分这两种方案在预测性能方面并无显著差异。

可以通过分箱绘制诊断图来检验假设（3.1），方法与 Henzi

等人 [13, 在分布回归情形下的图 2] 类似。预测性能应在与 $(y_i,\ \boldsymbol{x}_i)_{i=1}^n$互不重叠的数据测试集上评估，也就是说，使用在估计过程中完全未用过的数据。保序再校准可确保样本内自动校准；在独立同分布假设下，样本外也将近似满足自动校准。可按照 Gneiting-Resin [12], 的建议，利用 CORP（一致、最优分箱、可复现且采用 PAV）的均值可靠性图诊断样本外自动校准；预测性能则可使用常见的平方误差损失函数或偏差损失函数进行比较。

3.2 边界处的过拟合

保序再校准存在一个小问题，即它往往会在秩的下边界和上边界处过拟合 $\hat{\pi}(\boldsymbol{x}_1) < \ldots < \hat{\pi}(\boldsymbol{x}_n)$。例如，如果 y_n 是投资组合中最大的观测值（由于排序 $\hat{\pi}$ 是由响应数据驱动选定的，这种情况并非不可能），那么我们会估计 $\widehat{\mu}_{i_K} = y_n$，其中 $K = K((y_i,\ \hat{\pi}(\boldsymbol{x}_i))_{i=1}^n)$。这种做法往往会对（最小和最大）观测值过拟合，因为这类极端值/估计值无法通过样本外数据验证。因此，我们会直观地分析估计值 $\widehat{\boldsymbol{\mu}}$ 中的最大值和最小值，并可能手动将最小的块 $\mathcal{I}_1$ 与次小的块 $\mathcal{I}_2$ 合并（合并后的块上得到估计值 (2.4)）。更严格地说，可以用样本外数据对这种合并进行交叉验证，但我们不这样做。我们将在下文的图5中再次讨论这一点，届时会合并估计值最大的两个块。

3.3 解释

在 (2.3) 中，我们引入了复杂度数 $K((y_i,\ \hat{\pi}(\boldsymbol{x}_i))_{i=1}^n)$，它统计在保序再校准步骤中通过保序回归 (2.2) 得到的 $\widehat{\boldsymbol{\mu}}$ 中不同取值的数量。这个复杂度数 $K((y_i,\ \hat{\pi}(\boldsymbol{x}_i))_{i=1}^n)$ 可用于评估模型的信息含量，换言之，也就是有多少信号可以由数据解释。定理 2.3 表明，信噪比越低，我们预期保序回归的复杂度数也越低。显然，在定理 2.3 中，我们假设观测值的排序是正确的；但由于必须对 π 进行估计，这一假设只能近似满足。一般而言，如果样本量较大，并且使用灵活的回归模型对 π 进行建模，那么可以合理地认为这一结论在定性上仍然成立。然而，对于复杂的（算法型）回归模型，我们需要确保避免样本内过拟合；通常可通过使用（独立的）验证数据或进行交叉验证分析来控制。

非寿险中的典型理赔数据具有低信噪比。对于理赔频率而言，低信噪比是因为索赔并非十分频繁的事件。例如，在汽车保险中，年度理赔频率为 5% 至 10%，也就是说，在一个日历年内，每 10 名（或 20 名）驾驶员中只有 1 人会发生索赔。索赔金额也具有低信噪比，因为它们通常受随机性的强烈影响，而投保人信息所能提供的解释性部分相对有限。因此，我们通常认为，无论是理赔频率还是理赔金额建模，其复杂度数都较低 $K((y_i,\ \hat{\pi}(\boldsymbol{x}_i))_{i=1}^n)$。

当复杂度数 $K = K((y_i, \hat{\pi}(\boldsymbol{x}_i))_{i=1}^n)$较小到中等时，回归函数$\widehat{\boldsymbol{\mu}}$ 会通过保序再校准步骤变得易于解释。为此，我们将自动校准的回归函数 $\hat{\mu}$ 从集合 $\{\hat{\pi}(\boldsymbol{x}_1) \ldots, \hat{\pi}(\boldsymbol{x}_n)\}$ 扩展到整个协变量空间 $\mathcal{X}$

通过定义一个阶梯函数

$$\widehat{\mu}(\boldsymbol{x}) = \widehat{\mu}_{i_k}, \quad \text{if} \quad \widehat{\pi}(\boldsymbol{x}_{i_k}) \leq \widehat{\pi}(\boldsymbol{x}) < \widehat{\pi}(\boldsymbol{x}_{i_{k+1}}),$$

对于所有 $\boldsymbol{x} \in \mathcal{X}$，其中 $0 = i_0 < i_1 < \cdots < i_K = n$ 是式（2.3）中定义的保序回归的切分点。图1展示了这种阶梯函数插值，它不同于通常用于预测的插值方案。我们将原始协变量空间 $\mathcal{X}$ 划分为$\mathcal{X}_1, \ldots, \mathcal{X}_K$，

$$\mathcal{X}_k = \{\boldsymbol{x} \in \mathcal{X} : \widehat{\mu}(\boldsymbol{x}) = \widehat{\mu}_{i_k}\}, \quad k = 1, \ldots, K. \tag{3.2}$$

图4展示了如何通过对 $\mathcal{X}$ 进行划分，深入了解模型中的协变量-响应关系。这一过程与回归树和提升树有些相似，它们都依赖于对协变量空间 $\mathcal{X}$的划分。在第4节的案例研究中，我们展示了使用式（3.2）定义的划分来理解协变量-响应关系的另外两种方式。首先，图 7分析了各个协变量对价格群组的影响；其次，图9概述了所选价格群组对应的整个协变量空间。

4 瑞典摩托车数据

我们考虑对瑞典摩托车数据中的理赔金额进行建模。该数据最初发表于 Ohlsson–Johansson 的教科书中 [26] ，W¨uthrich– Merz [31]也对其进行了研究。[1] 这组数据涵盖瑞典的摩托车综合保险。该保险产品承保摩托车发生碰撞以外的损失或损坏，例如由盗窃、火灾或故意破坏造成的损失。数据包含按特征（协变量）组合汇总的 1994–1998日历年索赔记录。62,036 个不同协变量中共有 683 起索赔，因此索赔非常稀疏。我们完全采用 [31, 清单13.3],中所述的数据预处理方法，预处理后数据的一个摘录见清单1；关于各协变量的说明，请参见 [26, 第2节4] 和 [31, 第13节2]。我们的目标是为这 683 笔正索赔金额构建回归模型，并按照第3.3节所述，使用保序再校准实现自动校准和解释。

清单1：瑞典摩托车数据集摘录。

```
'data.frame':   62036 obs. of  9 variables:
 $ OwnerAge   : num  18 18 18 18 18 18 18 18 18 18 ...
 $ Gender     : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 1 1 1 1 1 ...
 $ Area       : Factor w/ 7 levels "Zone 1","Zone 2",..: 1 1 1 1 2 2 2 3 3 3 ...
 $ RiskClass  : int  1 2 3 3 3 1 1 3 1 1 ...
 $ VehAge     : num  8 11 9 9 11 12 2 4 4 6 6 ...
 $ BonusClass : int  2 2 3 4 1 1 2 1 1 2 ...
 $ Exposure   : num  1 0.778 0.499 0.501 0.929 ...
 $ ClaimNb    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ ClaimAmount: int  0 0 0 0 0 0 0 0 0 0 ...
```

[1]瑞典摩托车数据集可通过 R 软件包 CASdatasets 获取 [10]。

4.1 保序再校准与二元回归树的比较

我们首先只考虑风险类别和车辆年龄这两个协变量分量。由于由此得到的协变量空间 $\mathcal{X} = \{$（风险类别，车辆年龄）$\} \subset \mathbb{R}^2$ 是二维的，因此我们可以通过图示说明保序再校准方法与二元回归树（作为一种竞争模型）在解释方面的差异。在第4.2节中，我们将考虑所有可用协变量。

我们对这 683 项索赔拟合了一个深度前馈神经网络（FFNN）回归模型。我们选择深度为 3 的网络架构，三个隐藏层分别包含 (20，15，10) 个神经元，隐藏层采用双曲正切激活函数，输出层采用对数连接函数。输入维度为 2，因此该前馈神经网络的网络参数维度为 546；关于前馈神经网络的更详细讨论，请参见 [31, Chapter 7], ，尤其是该参考文献中的清单 7.1–7.3。我们使用伽马偏差损失拟合此模型，见 [31,第5.3节。7]和注2.2；采用随机梯度下降的 Nadam 版本，并在占全部数据 20% 的验证集上执行提前停止。表1第 (1a) 行给出了拟合的前馈神经网络回归模型的表现，该行称为伽马前馈神经网络。我们将其与第 (0) 行的零模型（经验均值）进行比较，后者不考虑任何协变量。[2] 我们观察到伽马偏差损失和均方根误差（RMSE）均有所降低，这证明采用回归模型是合理的；请注意，这些是样本内指标，但我们通过提前停止来防止网络出现样本内过拟合。这里的难点在于，只有 683 项索赔，因此无法进行合理的样本外分析。表1最后一列名为“平均值”，比较了前馈神经网络的平均索赔估计与经验均值；我们观察到前馈神经网络预测存在轻微的正偏差，即 24，$932 > 24$，641。

		伽马偏差	RMSE	平均值
(0)	零模型	2.085	35,311	24,641
(1a)	伽马前馈神经网络	1.704	32,562	24,932
(1b)	重新校准后的伽马前馈神经网络	1.640	32,005	24,641
(2)	二元回归树	1.761	32,706	24,641

表1：瑞典摩托车示例中仅将风险类别和车辆年龄作为协变量时的损失数值。

在下一步中，我们使用前馈神经网络的估计值作为秩 $\widehat{\pi}(\boldsymbol{x}_i)$，分别用于对索赔 y_i 和协变量 $\boldsymbol{x}_i$ 排序。随后，我们对这些秩和索赔应用非参数保序再校准步骤 (2.2)。瑞典摩托车索赔数据按可用的协变量组合进行了聚合，683 个正索赔来自 656 种不同的协变量组合 $\boldsymbol{x}_i$。因此，我们需要使用 (2.2) 的加权版本，其中 $w_i \in \mathbb{N}$对应于协变量 $\boldsymbol{x}_i$下观测到的索赔数量，而 y_i 对应于 $\boldsymbol{x}_i$ 上的平均观测索赔金额。[3] 我们使用 R 软件包 monotone [7] ，它提供了

[2]在伽马零模型中，也就是说，假设响应变量独立同分布且服从伽马分布，我们得到均值的最大似然估计等于观测值的经验均值；这通常在指数离散族中成立。[3]由于本例仅考虑风险类别和车辆年龄这两个协变量分量，我们进一步按这些协变量组合汇总索赔。这为伽马回归模型提供了充分统计量，因此我们只需相应地调整权重 w_i 。对于连续回归函数而言，这是一种巧妙的方式，可以避免处理并列值（并假设不同协变量组合内的汇总

PAV 算法的快速实现。数值结果列于表1的第 (1b) 行。经过保序再校准后，平均损失略有下降。这是预期之中的，因为保序回归针对任何 Bregman 损失函数优化样本内损失，参见注2.2。表1的最后一列验证了此时全局平衡性质 (2.6) 成立。

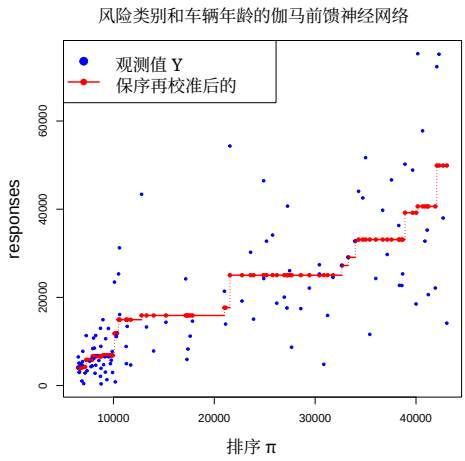


图3：在瑞典摩托车示例中，仅使用风险类别和车辆年龄作为协变量进行保序再校准，得到复杂度数 $K((y_i,\ \widehat{\pi}(\boldsymbol{x}_i))_{i=1}^n) = 18$。

图3展示了对伽马前馈神经网络给出的排序 $\widehat{\pi}(\boldsymbol{x}_i))_{i=1}^n$ 进行保序再校准后得到的阶梯函数（红色）；图中还补充显示了观测金额 y_i （蓝色）。得到的复杂度数为 $K = K((y_i,\ \widehat{\pi}(\boldsymbol{x}_i))_{i=1}^n) = 18$，也就是说，在此示例中，条件期望赔付金额可由18个不同的估计值 $\widehat{\mu}_{i_k} \in \mathbb{R}$, $k = 1, \ldots,$ $K = 18$表示；前馈神经网络回归函数使用了 $6 \cdot 21 = 126$个不同的值（秩），这对应于可用协变量值（风险类别，车辆年龄）的基数 $\in \mathcal{X}$。

对前馈神经网络的秩 $\widehat{\pi}(\boldsymbol{x}) = \widehat{\pi}$(风险类别，车辆年龄)进行保序再校准后，得到协变量空间的一个划分 $\mathcal{X}_1$，……，$\mathcal{X}_{18}$，其定义见式（3.2）。我们将此划分与采用二元分裂回归树方法得到的划分进行比较。我们使用10折交叉验证来确定最优树大小。在此例中，最优树只有3个分裂，且都涉及变量车辆年龄。该最优树产生的损失显示在表1第（2）行，据此我们得出结论：在此处，回归树方法并不完全具有竞争力。更有意思的是，图4展示了两种方法得到的协变量空间 $\mathcal{X} = \{$(风险类别，车辆年龄)$\}$ 划分。右侧图展示了回归树的三个分裂（均针对车辆年龄）。从左侧的保序再校准方法中，我们可以看出，良好的回归模型应具有对角结构，这表明这两个协变量以一种非平凡的方式相互作用，而在此情况下，二元分裂回归树无法捕捉这种关系。

在计算上是可行的）。

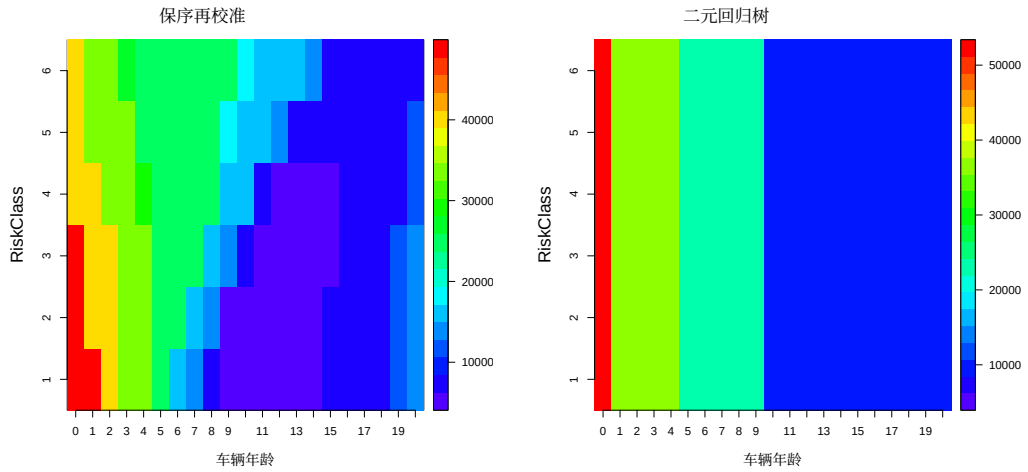


图4：（左）保序再校准和（右）二元回归树，两者均仅使用风险类别和车辆年龄作为协变量；两幅图的颜色刻度相同。

4.2 考虑所有协变量

现在我们考虑所有可用的协变量分量，见清单1第2至7行。我们首先用这些数据拟合一个前馈神经网络。具体做法与前一个例子完全相同，唯一的区别是，考虑所有可用信息时，输入维度从2变为6。我们将（有序的）区域代码转换为实数，并且由于区域代码5至7对应的索赔稀少，也将它们合并，并将这个新变量称为区域。此时，前馈神经网络的网络参数维度为626。我们使用随机梯度下降拟合该网络，并基于验证损失分析进行提前停止。结果见表2第（2a）行。

		伽马偏差	RMSE	平均值
(0)	零模型	2.085	35,311	24,641
(1a)	伽马广义线性模型	1.717	32,562	25,105
(1b)	使用 $K = 24$ 重新校准的伽马广义线性模型	1.641	31,578	24,641
(2a)	伽马前馈神经网络	1.496	29,673	24,526
(2b)	使用 $K = 22$ 重新校准的伽马前馈神经网络	1.452	28,806	24,641
(2c)	伽马前馈神经网络树调整，使用 4 个分箱（随机种子 1）	1.508	29,371	24,641
(2d)	伽马前馈神经网络树调整，使用 8 个分箱（随机种子 2）	1.466	27,942	24,641

表 2：基于所有可用协变量的瑞典摩托车示例中的损失。

我们将拟合得到的前馈神经网络回归模型与零模型（经验均值）以及伽马广义线性模型（广义线性模型）进行比较。该伽马广义线性模型与 [31,表5.13]中的模型 Gamma GLM1 完全相同。下面对表2的结果做几点说明。首先，在第 (0) 至 (2a) 行中，前馈神经网络的伽马偏差损失和均方根误差均为三个模型中最小。因此，在这三种模型选择中，伽马前馈神经网络对数据的适应效果最佳（拟合前馈神经网络时，我们采用了提前停止）。值得注意的是，伽马广义线性模型和前馈神经网络都不满足全局平衡性质 (2.6)，见表2最后一列。采用提前停止拟合的随机梯度下降模型通常不满足全局平衡性质，而伽马

广义线性模型不满足全局平衡性质，是因为这里我们使用的是对数连接函数，而不是伽马广义线性模型的规范链接函数。

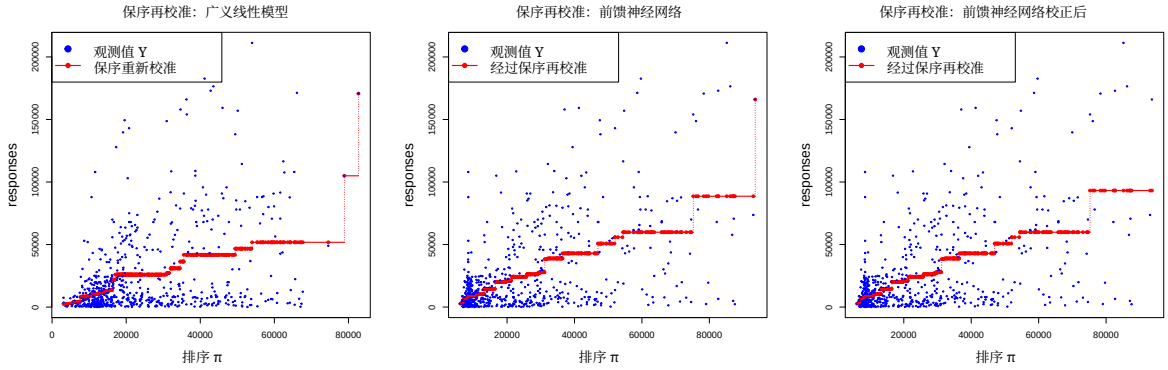


图5：在瑞典摩托车示例中，使用所有协变量对回归模型进行保序再校准：复杂度数为 $K((y_i, \hat{\pi}(\boldsymbol{x}_i))_{i=1}^n) = 24$ 的伽马广义线性模型（左），复杂度数为 $K((y_i, \hat{\pi}(\boldsymbol{x}_i))_{i=1}^n) = 23$ 的伽马前馈神经网络（中），以及经过过拟合校正的模型（右）。

接下来，我们使用前馈神经网络预测作为秩 $\hat{\pi}(\boldsymbol{x}_i)$，据此排列响应变量和协变量，并将索赔标记为 y_i，使得 $\hat{\pi}(\boldsymbol{x}_1) < \ldots < \hat{\pi}(\boldsymbol{x}_n)$。这些数据中不存在并列情况，因此我们得到 $n = 656$ 个两两不同的值。保序再校准的结果见图5（中）。复杂度数为 $K = K((y_i, \hat{\pi}(\boldsymbol{x}_i))_{i=1}^n) = 23$，因此，整个回归问题由23个不同的值 $\widehat{\mu}_{i_k}$, $k = 1, \ldots, K$ 编码。根据这张图，最大的值 $\widehat{\mu}_{i_K}$ 似乎对相应观测值过拟合，因为这个估计由单个观测值 y_n 决定，且大于前一个块 $\mathcal{I}_{K-1}$ 上的加权块均值 $\widehat{\mu}_{i_{K-1}}$；参见第3.2节。为此，我们手动合并最后两个块 $\mathcal{I}_{K-1}$ 和 $\mathcal{I}_K$。这样一来，我们便在合并后的块上得到新的估计值 (2.4)，并将复杂度数减少1，降至 $K = 22$。最终的保序再校准结果见图5（右），经验损失列于表2第 (2b) 行。重要的是，这种保序再校准后的回归在样本内自动校准 (2.5)，因此满足全局平衡性质，这一点可由表2的最后一列验证。

我们对表2中伽马广义线性模型得到的秩执行相同的保序再校准。我们观察到，保序再校准步骤显著降低了伽马广义线性模型的平均损失，并使复杂度数达到 $K = 24$，另请参见图5（左）。

我们将保序再校准与 Lindholm 等人近期提出的方法进行比较，后者也能在样本内实现自动校准。保序回归将指标集 $\mathcal{I} = \{1, \ldots, n\}$划分为互不相交的块 $\mathcal{I}_1, \ldots, \mathcal{I}_K$，在这些块上，估计回归函数为常数。这也可以通过对（排序后的）协变量 $\{\hat{\pi}(\boldsymbol{x}_i); 1 \leq i \leq n\}$ 及其对应的响应变量 y_i 应用二元回归树算法来实现；参见 Lindholm 等人的第 2.3.2 节 [21]。我们将后一种方法称为树分箱方法。树分箱方法与保序再校准方法有两个主要区别。首先，通常情况下，树分箱方法得到的回归函数，其排序与第一步回归得到 $\hat{\pi}(\boldsymbol{x}_i)$ 时所用的排序并不相同。其次，在保序回归方法中，复杂度数 $K((y_i, \hat{\pi}(\boldsymbol{x}_i))_{i=1}^n)$ 是自然确定的，即保序回归 (2.2) 会自动

提取响应变量 $\boldsymbol{y}$ 中包含的信息量；根据定理 2.3，通常而言，随着信噪比提高，这一信息量也会增加。相反，在树分箱方法中，我们需要确定最优分箱数（叶节点数），例如通过 k 折交叉验证。得到的分箱数取决于叶节点最小大小和交叉验证折数这两个超参数，也取决于交叉验证时实例的随机划分。我们发现，分箱数对调参选择较为敏感，因此，与保序再校准不同，最终的划分可能受到主观选择和随机性的影响。

对于表2中树分箱方法的结果，我们选择了 $k = 10$ 折，最小叶节点大小为 10；而第 (2c)–(2d) 行结果的差异仅在于伪样本的随机划分不同。使用第一个随机种子得到 4 个分箱，使用第二个随机种子得到 8 个分箱；我们观察到，这两个模型在伽马偏差损失和均方根误差方面存在显著差异。图 6 展示了保序再校准以及采用 8 个分箱的树分箱方法，分别对应表2中的第 (2b) 行和第 (2d) 行。从图中可以得出结论，树分箱方法不一定能保留由 $\widehat{\pi}(\boldsymbol{x}_i)$ 导出的排序，因为得到的阶梯函数（蓝色）并非单调递增。我们建议采用保序再校准来实现自动校准，因为它能在第一个估计步骤中保留回归模型的单调性，而且不存在可能产生影响的调优参数。

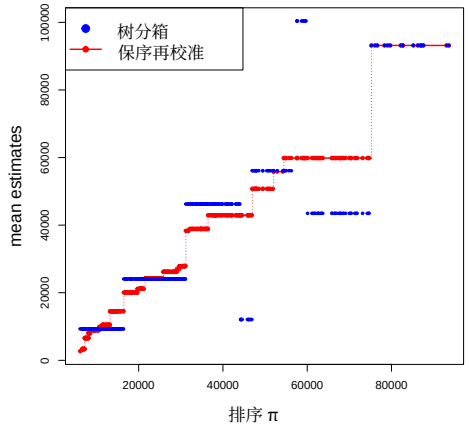


图 6：树分箱与保序再校准；阶梯函数分别对应表2中的公式 (2d) 和 (2b)，其中公式 (2d) 使用 8 个分箱，公式 (2b) 的复杂度数为 $K = 22$ 。

在图 7 中，我们展示了将保序再校准的估计值 $\widehat{\boldsymbol{\mu}}$ 投影到相应协变量值后得到的边际图，也就是说，这是结果协变量空间划分 (3.2) 的边际视图。对于较低的复杂度数 $K((y_i, \widehat{\pi}(\boldsymbol{x}_i))_{i=1}^n)$，这种视图很容易解释。我们可以看到，在车主年龄、区域、风险类别和车辆年龄的不同协变量水平下，颜色分布存在明显差异。这表明这些变量对于解释索赔金额很重要，但需要注意，这种边际视图忽略了潜在的交互作用。对于性别变量，我们无法得出任何结论，因为性别比例不平衡过于严重。奖励等级的解释则不那么直观。事实上，根据伽马广义线性模型，我们知道奖励等级并不显著，见 [31, 表 5.13]。这是因为奖励等级与碰撞索赔有关，而我们的数据研究的是不包括碰撞索赔的综合保险。图 8 展示了保序

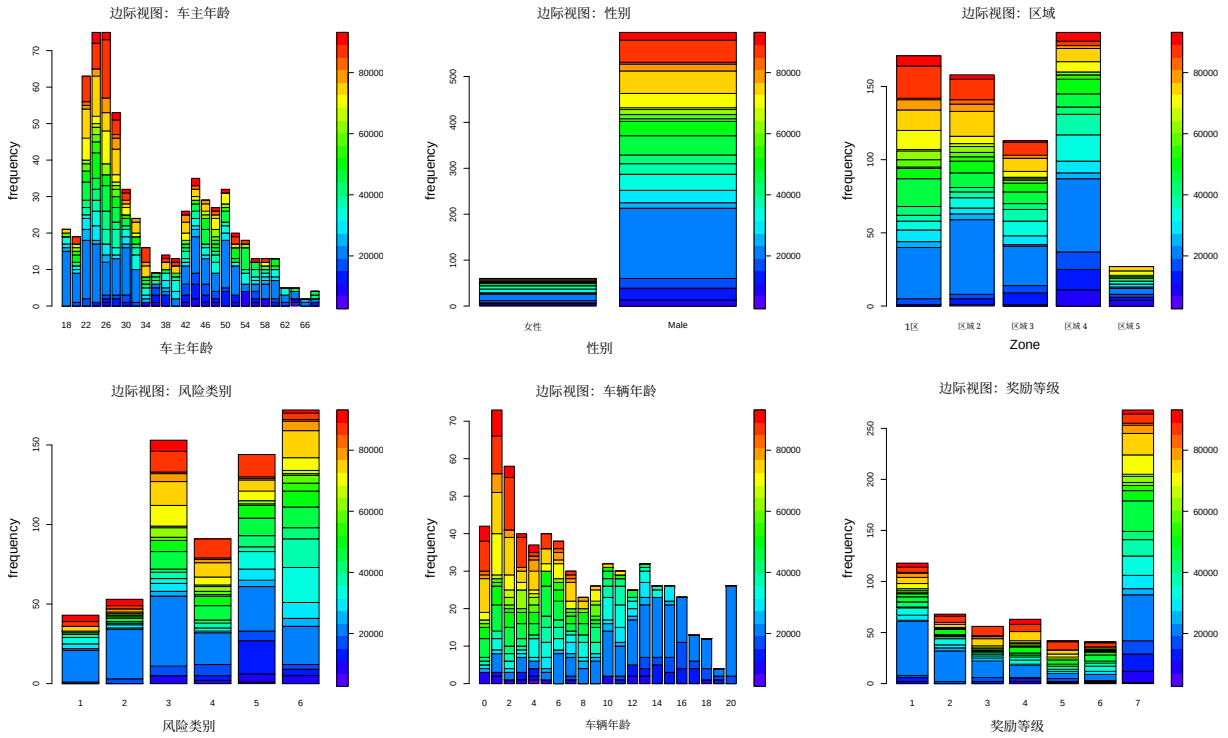


图 7：表2中经等渗重校准的伽马前馈神经网络模型的边际视图。该模型考虑了 6 个协变量分量：车主年龄、性别、区域、风险类别、车辆年龄和奖励等级。

重新校准后的伽马前馈神经网络（左）和伽马广义线性模型（右）在协变量奖励等级上的边际视图。如前所述，奖励等级在伽马广义线性模型中并不显著；从图中看来，两个模型在不同等级上的颜色分布确实相当相似。

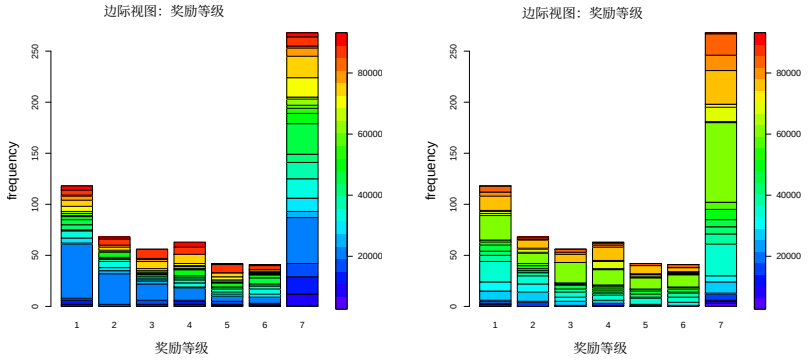


图 8：经等渗重校准的伽马前馈神经网络模型（左）和经等渗重校准的伽马广义线性模型（右）中协变量分量奖励等级的边际视图。

显然，车辆年龄是最重要的变量，呈现出新摩托车的索赔金额更高这一现象。不同区域的索赔金额分布存在显著差异，其中1区是瑞典最大的三个城市所在区域，通常会出现更多大额索赔。风险类别对应摩托车的大小，并与车主年龄、车辆年龄和区域相互作用，因此更难以解释，因为这些变量之间存在相关的交互作用。

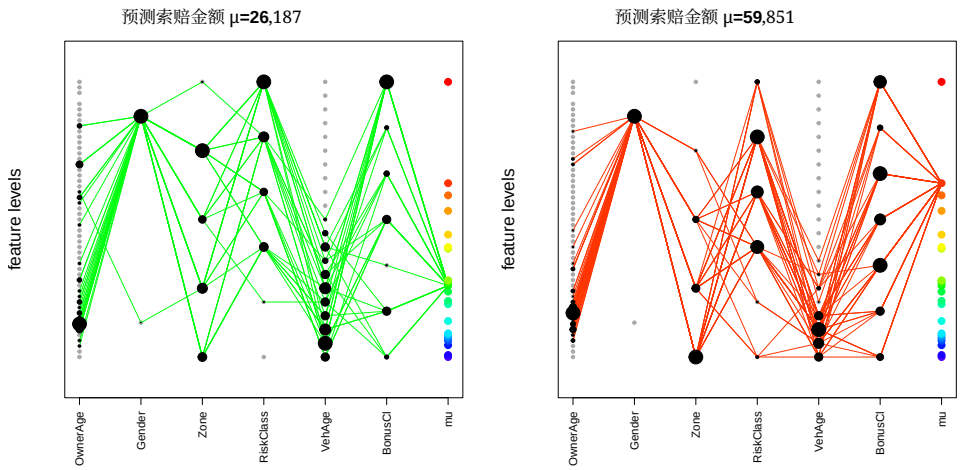


图9：针对 $k = 12,21$的两个选定值进行保序再校准时，协变量空间的划分 $(\mathcal{X}_k)_{k=1,\ldots,K}$ $\mathcal{X}$。

图9展示了针对保序再校准 $(\widehat{\mu}_{i_k})_{k=1,\ldots,K}$，六维协变量空间的划分 $(\mathcal{X}_k)_{k=1,\ldots,K}$ 在两个选定的 k 值下的示例。线条连接数据中观察到的、具有给定 $\widehat{\mu}_{i_k}$ 值的所有协变量分量 $\boldsymbol{x}$ $(\boldsymbol{x}_i)_{1\leq i\leq n}$，黑点的大小则表示某个协变量水平被观察到的频次。例如，右侧的图对应第二高的索赔预测值 $\widehat{\mu}_{i_{K-1}} = 59,851$。对于这一期望响应水平，车主年龄相对较小（约25岁）；所有人都是男性，且大多居住在1区（瑞典最大的三个城市），拥有风险类别较高、车辆年龄较低的摩托车。对协变量空间 $\mathcal{X}$ 的其他部分 $\mathcal{X}_k$ 也可以得出类似结论。因此，较低的复杂度数 $K((y_i,\ \widehat{\pi}(\boldsymbol{x}_i))_{i=1}^n)$ 有助于解释回归模型。

5 结论

我们处理了两个问题。首先，我们对一个已拟合的（初始）回归模型的秩进行保序再校准，从而确保回归模型满足自动校准性质。这种保序再校准不涉及任何超参数，只是假设初始回归模型给出的秩是（近似）正确的。保序回归具有这样一项性质：在低信噪比问题中，所得（非参数）回归函数的复杂度较低。借助这一性质，我们证明了这种方法能够得到可解释回归函数，因为复杂度较低等价于协变量空间的划分较为粗略。在保险定价问题中，这一点尤其有用，因为保险理赔数据通常具有较低的信噪比。于是，我们可以先对这些数据拟合一个复杂的（算法型）模型，随后再使用保序再校准对初始回归函数进行自动校准；这也会显著简化回归函数。

参考文献

[1] Ayer, M., Brunk, H.D., Ewing, G.M., Reid, W.T., Silverman, E. (1955). 基于不完整信息抽样的经验分布函数。《数学统计年鉴》26, 641-647.[2] Balabdaoui, F., Durot, C., Jankowski, H. (2019). 单调单指标模型中的最小二乘估计。《伯努利》25, 3276-3310.[3] Barlow, R.E., Bartholomew, D.J., Brenner, J.M., Brunk, H.D. (1972). 《序约束下的统计推断》。Wiley.[4] Barlow, R.E., Brunk, H.D. (1972). 保序回归问题及其对偶。《美国统计学会杂志》67/337, 140-147.[5] Best, M.J., Chakravarti, N. (1990). 保序回归的活动集算法：一个统一框架。《数学规划》47 425-439.[6] Brunk, H.D., Ewing, G.M., Utz, W.R. (1957). 单调函数某些类中的积分最小化。《太平洋数学杂志》7, 833-847.[7] Busing, F.M.T.A. (2022). 单调回归：一种简单快速的 $O(n)$ 池相邻违背者算法实现。《统计软件杂志》102, 代码片段 1.[8] Denuit, M., Charpentier, A., Trufin, J. (2021). 机器学习中的保险定价：自动校准与Tweedie支配。《保险：数学与经济学》101/B, 485-497.[9] Dimitriadis, T., D¨umbgen, L., Henzi, A., Puke, M., Ziegel, J. (2022). 二元结果预测的诚实校准评估。arXiv:2203.04065.[10] Dutang, C., Charpentier, A. (2018). CASdatasets R软件包指南。参考手册。版本 1.0-8，打包于 2018-05-20.[11] Gneiting, T. (2011). 制作和评估点预测。《美国统计学会杂志》106/494, 746-762.[12] Gneiting, T. (2021). 回归诊断与预测评估相结合：条件校准、可靠性图和决定系数。arXiv:2108.03210.[13] Henzi, A., Kleger, G.-R., Ziegel, J.F. (2021). 分布（单指标）模型。《美国统计学会杂志》，即将刊出.[14] Henzi, A., Ziegel, J.F., Gneiting, T. (2021). 保序分布回归。《英国皇家统计学会杂志：B辑》85, 963-993.[15] Karush, W. (1939). 《带不等式侧约束的多变量函数极小值》。理学硕士论文。芝加哥大学数学系。[16] Kingma, D.P., Welling, M. (2019). 变分自编码器简介。《机器学习基础与趋势》12/4, 307-392.[17] Kr¨uger, F., Ziegel, J.F. (2021). 预测支配的一般条件。《商业与经济统计杂志》39/4, 972-983.[18] Kruskal, J.B. (1964). 非度量多维尺度分析。Psychometrica 29, 115-129.[19] Kuhn, H.W., Tucker, A.W. (1951). 非线性规划。第二届伯克利研讨会论文集。加利福尼亚大学出版社，481-492.[20] Leeuw, de J., Hornik, K., Mair, P. (2009). R中的保序优化：池相邻违背者算法（池相邻违背者算法）和活动集方法。《统计软件杂志》32/5, 1-24。

[21] Lindholm, M., Lindskog, F., Palmquist, J. (2022). 局部偏差调整、按期限加权的概率，以及费率单元的自动构建。*SSRN*手稿ID 4256876.[22] Lundberg, S.M., Lee, S.-I. (2017). 一种解释模型预测的统一方法。载于：神经信息处理系统进展**30**。Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R.（编）。Curran Associates，4765-4774。[23] Menon, A.K., Jiang, X., Vembu, S., Elkan, C., Ohno-Machado, L.（2012）。利用排序损失预测准确概率。*ICML'12*：第*29*届国际机器学习会议论文集，659-666。[24] Miles, R.E.（1959）。通过加权均值将有限实数集合完全合并为若干块。*Biometrika* **46**，317-327。[25] Patton, A.J.（2011）。使用不完善的波动率代理指标进行波动率预测比较。计量经济学杂志**160**，246-256。[26] Ohlsson, E., Johansson, B.（2010）。基于广义线性模型的非寿险定价。施普林格。[27] Savage, L.J.（1971）。个人概率与期望的可报告性。《美国统计学会杂志》**66/336**，783-810。[28] Schervish, M.J.（1989）。比较概率评估者的一般方法。统计年鉴**17/4**，1856-1879。[29] Tasche, D.（2021）。充分校准。统计学：理论与应用期刊**55/6**，1356-1386。[30] W¨uthrich M.V.（2022）。自动校准条件下基于基尼指数的模型选择。欧洲精算学期刊，即将出版。[31] W¨uthrich, M.V., Merz, M.（2023）。精算学习及其应用的统计学基础。Springer Actuarial。[32] Zadrozny, B., Elkan, C.（2002）。将分类器得分转换为准确的多类别概率估计。第八届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，694-699。

A 附录

A.1 相邻违序算法

最小化问题（2.2）是一个带有线性侧约束的二次优化问题，可以使用 Karush–Kuhn–Tucker (KKT) 方法求解 [15, 19]。因此，我们考虑其拉格朗日函数

$$L(\boldsymbol{\mu}, \boldsymbol{\eta}) = (\boldsymbol{y} - \boldsymbol{\mu})^\top W(\boldsymbol{y} - \boldsymbol{\mu}) - \boldsymbol{\eta}^\top A\boldsymbol{\mu},$$

其中拉格朗日乘子为 $\boldsymbol{\eta} \in \mathbb{R}^{n-1}$。KKT 条件如下：

$$\begin{aligned}
\mathbf{0} &= \nabla_{\boldsymbol{\mu}} L(\boldsymbol{\mu}, \boldsymbol{\eta}) = -W(\boldsymbol{y} - \boldsymbol{\mu}) - A^\top \boldsymbol{\eta}, && \text{(A.1)}\\
\mathbf{0} &\geq \nabla_{\boldsymbol{\eta}} L(\boldsymbol{\mu}, \boldsymbol{\eta}) = -A\boldsymbol{\mu}, && \text{(A.2)}\\
\mathbf{0} &\leq \boldsymbol{\eta}, && \text{(A.3)}\\
\mathbf{0} &= (\eta_1(\mu_1 - \mu_2), \ldots, \eta_{n-1}(\mu_{n-1} - \mu_n))^\top. && \text{(A.4)}
\end{aligned}$$

这些 KKT 条件 (A.1)-(A.4) 的解给出了保序估计 $\widehat{\boldsymbol{\mu}}$。该解可以通过 PAV 算法求得。其主要思路是比较原始估计 $(\widetilde{\mu}_i)_i$。如果存在满足 $\widetilde{\mu}_i > \widetilde{\mu}_{i+1}$ 的相邻对，则违反了单调性约束。此类相邻对会递归地合并（汇集）为一个具有相同估计值的块；依次汇集违反单调性约束的相邻对和块，即可得到 PAV 算法。

Pool Adjacent Violators (PAV) Algorithm

---

(0) 初始化 t　该算法 $\widehat{\boldsymbol{\mu}}^{(0)} = \boldsymbol{y}$ 并定义块 $\mathcal{I}_k^{(0)} = \{k\}$，其中 $k = 1$，……，$K^{(0)} = n$.

(1) 对 $t \geq 0$ 进行迭代：

(a) 如果 $\widehat{\boldsymbol{\mu}}^{(t)}$ 满足 KKT 条件 (A.2)，则转到第 (2) 项，否则转到下一步 (1b)。(b) 选择一个指标 $i = 1, \ldots, n$，满足 $\widehat{\mu}_i^{(t)} > \widehat{\mu}_{i+1}^{(t)}$，将包含 $i \in \mathcal{I}_k^{(t)}$ 和 $i+1 \in \mathcal{I}_{k+1}^{(t)}$ 的两个相邻区块合并，并保持其他所有块不变。由此得到新的块 $\mathcal{I}_k^{(t+1)}$，满足 $k = 1, \ldots,$ $K^{(t+1)} = K^{(t)} - 1$。

(c) 在每个块 $k = 1, \ldots, K^{(t+1)}$ 上设定，并对所有指标 $i \in \mathcal{I}_k^{(t+1)}$，设定新的估计值

$$\widehat{\mu}_i^{(t+1)} = \frac{1}{\sum_{j \in \mathcal{I}_k^{(t+1)}} w_j} \sum_{j \in \mathcal{I}_k^{(t+1)}} w_j y_j. \tag{A.5}$$

(d) 增大 $t \mapsto t + 1$。

(2) 设定保序回归估计 $\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}^{(t)}$，若相邻区块 $\mathcal{I}_k^{(t)}$ 和 $\mathcal{I}_{k+1}^{(t)}$ 上的估计值 $\widehat{\mu}_i$ 相同，则合并这些区块。返回 $\mathcal{I}$ 的结果划分，记为 $(\mathcal{I}_k)_{k=1,\ldots,K}$ 和 $\widehat{\boldsymbol{\mu}}$。

备注 A.1 (PAV 算法解释**)**

(0) 我们从无约束最优解开始，并设定 $\boldsymbol{\eta}^{(0)} = 0$ 以确保满足 KKT 条件 (A.1)、(A.3) 和 (A.4)，因此，只有单调性 (A.2) 不一定得到满足。

(1a) 我们找出一对违反单调性约束 (A.2) 的 $\widehat{\mu}_i^{(t)} > \widehat{\mu}_{i+1}^{(t)}$ 。这对元素必须分别属于两个相邻区块 $\mathcal{I}_k^{(t)}$ 和 $\mathcal{I}_{k+1}^{(t)}$，因为每个区块内的估计值都是常数 (2.4)。我们将这两个相邻区块合并为 $\mathcal{I}_k^{(t+1)} = \mathcal{I}_k^{(t)} \cup \mathcal{I}_{k+1}^{(t)}$，从而使区块数 $K^{(t)}$ 减少 1。

(1b) 我们在每个区块上设定常数估计值 (A.5)，使其满足区块内的单调性约束 (A.2)；同时，条件 (A.4) 在该区块内也自然成立。通过相应地调整拉格朗日参数 $\boldsymbol{\eta}^{(t)} \mapsto \boldsymbol{\eta}^{(t+1)} \geq \mathbf{0}$，以考虑 (A.1) 中均值估计值 (A.5) 的变化，即可满足条件 (A.1) 和 (A.3)。

(1c) 对于大小为 n 的样本，此算法至多可迭代 $n-1$ 次，因此算法将会终止。

(2) 由于对于 $i \in \mathcal{I}_k^{(t)}$ 和 $i+1 \in \mathcal{I}_{k+1}^{(t)}$，我们有不等式 $\widehat{\mu}_i^{(t)} \leq \widehat{\mu}_{i+1}^{(t)}$，最后一步是通过合并相应估计值不满足严格不等式的块，确保所得的块是极大的。

A.2 定理 2.3 的证明

定理 2.3 的证明。 对于给定的响应变量 $\boldsymbol{y} = \boldsymbol{Y}_\sigma(\omega)$，(2.2) 的解给出了指标集 $\mathcal{I}$ 的划分 (2.3)，并在块 $\mathcal{I}_k$ 上得到经验加权平均值 (2.4)。由于块 $\mathcal{I}_k$ 已选为极大块，对于所有 $k = 1$、.....、$K(\boldsymbol{Y}) - 1$，这些经验加权平均值满足 $\widehat{\mu}_{i_k} < \widehat{\mu}_{i_{k+1}}$。现在我们来考察这些块在 PAV 算法中是如何构造的。假设当前处于第 $t \geq 0$ 次迭代，并且在这次 PAV 算法迭代中，我们合并相邻区块 $\mathcal{I}_k^{(t)}$ 和 $\mathcal{I}_{k+1}^{(t)}$，因为对于 $i \in \mathcal{I}_k^{(t)}$ 和 $i+1 \in \mathcal{I}_{k+1}^{(t)}$，有 $\widehat{\mu}_i^{(t)} > \widehat{\mu}_{i+1}^{(t)}$。我们来分析这个不等式

$$\frac{1}{\sum_{j \in \mathcal{I}_k^{(t)}} w_j} \sum_{j \in \mathcal{I}_k^{(t)}} w_j y_j = \widehat{\mu}_i^{(t)} > \widehat{\mu}_{i+1}^{(t)} = \frac{1}{\sum_{l \in \mathcal{I}_{k+1}^{(t)}} w_l} \sum_{l \in \mathcal{I}_{k+1}^{(t)}} w_l y_l.$$

我们使用位置-尺度结构 (2.4)，由此得到等价条件

$$\frac{1}{\sum_{j \in \mathcal{I}_k^{(t)}} w_j} \sum_{j \in \mathcal{I}_k^{(t)}} w_j \left(\mu_j + \sigma \epsilon_j(\omega)\right) > \frac{1}{\sum_{l \in \mathcal{I}_{k+1}^{(t)}} w_l} \sum_{l \in \mathcal{I}_{k+1}^{(t)}} w_l \left(\mu_l + \sigma \epsilon_l(\omega)\right).$$

由于对于任意指标 $j \in \mathcal{I}_k^{(t)}$ 和 $l \in \mathcal{I}_{k+1}^{(t)}$，我们都有 $\mu_j \leq \mu_l$，因此在 PAV 算法的第 t 次迭代中合并相邻区块 $\mathcal{I}_k^{(t)}$ 和 $\mathcal{I}_{k+1}^{(t)}$ 的前述条件可写为

$$\sigma \left[\frac{\sum_{j \in \mathcal{I}_k^{(t)}} w_j \epsilon_j(\omega)}{\sum_{j \in \mathcal{I}_k^{(t)}} w_j} - \frac{\sum_{l \in \mathcal{I}_{k+1}^{(t)}} w_l \epsilon_l(\omega)}{\sum_{l \in \mathcal{I}_{k+1}^{(t)}} w_l} \right] > \frac{\sum_{l \in \mathcal{I}_{k+1}^{(t)}} w_l \mu_l}{\sum_{l \in \mathcal{I}_{k+1}^{(t)}} w_l} - \frac{\sum_{j \in \mathcal{I}_k^{(t)}} w_j \mu_j}{\sum_{j \in \mathcal{I}_k^{(t)}} w_j} \geq 0. \tag{A.6}$$

重要的观测结果是，如果尺度参数 $\sigma > 0$满足这一条件，那么任何更大的尺度参数 $\sigma' > \sigma > 0$ 也满足这一条件（对 $\omega \in \Omega$ 中的每个点均成立）。因此，

任何在 σ 下发生的合并，在 $\sigma' > \sigma > 0$下也会发生。由于这对底层概率空间（Ω $\mathcal{F}$， $\mathbb{P}$）中的每个点都成立，因此可知 $\mathbb{E}[K(\boldsymbol{Y})]$ 随 $\sigma > 0$增大而递减。现在假设 $\boldsymbol{\epsilon}$ 的分布在 $\mathbb{R}^n$上具有全支撑。那么，事件 $A_\sigma = \{K(\boldsymbol{Y}_\sigma) = n\}$ 以正概率发生，即，

$$\begin{aligned} 0 &< \mathbb{P}[A_\sigma] = \mathbb{P}\left[K(\boldsymbol{Y}_\sigma) = n\right] \\ &= \mathbb{P}\left[Y_1 < Y_2 < \ldots < Y_n\right] \\ &= \mathbb{P}\left[\mu_1 + \sigma\epsilon_1 < \mu_2 + \sigma\epsilon_2 < \ldots < \mu_n + \sigma\epsilon_n\right]. \end{aligned}$$

考虑

$$\begin{aligned} A_\sigma &= \{\mu_1 + \sigma\epsilon_1 < \mu_2 + \sigma\epsilon_2 < \ldots < \mu_n + \sigma\epsilon_n\} \\ &= \bigcap_{k=2}^{n} \{\mu_{k-1} + \sigma\epsilon_{k-1} < \mu_k + \sigma\epsilon_k\} \\ &= \{\sigma(\epsilon_1 - \epsilon_2) < \mu_2 - \mu_1\} \cap \bigcap_{k=3}^{n} \{\mu_{k-1} + \sigma\epsilon_{k-1} < \mu_k + \sigma\epsilon_k\}. \end{aligned}$$

我们关注右侧的第一个事件。注意到 $\mu_2 - \mu_1 > 0$，因此

$$\left\{\boldsymbol{\epsilon} \in \mathbb{R}^n : \epsilon_1 - \epsilon_2 < \frac{\mu_2 - \mu_1}{\sigma}\right\}$$

描述了 $\mathbb{R}^n$ 中的一个开半空间，该半空间包含原点，其边界超平面为 ~，并且当 σ 减小时，该超平面会进一步远离原点。总之， $A_\sigma \subset \mathbb{R}^n$ 中 $\boldsymbol{\epsilon}$ 在 A_σ 内的取值集合是一个包含原点的非空开多面体，并且会随 σ 缩放，即 ~ ~A_σ =(1/σ)A_1。因此，由于 $\boldsymbol{\epsilon}$ 的分布具有全支撑，概率 $\mathbb{P}[A_\sigma]$ 会随 σ 严格递减。

□